Lucrările conferinței
Resurse lingvistice și instrumente pentru prelucrarea limbii române
București, 6–7 mai 2010

Volum apărut cu sprijinul Ministerului Educației, Cercetării și Tineretului, prin Centrul Național de Management Programe (CNMP), proiectul eDTLR – Dicționarul Tezaur al Limbii Române în format electronic (contract 91–013/18.09.2007)

ISSN 1843-911X

Lucrările conferinței
Resurse lingvistice și instrumente pentru prelucrarea limbii române
București, 6-7 mai 2010

Editori:
Adrian Iftene
Horia-Nicolai Teodorescu
Dan Cristea
Dan Tufiș

Organizatori:
Facultatea de Informatică,
Universitatea „Alexandru Ioan Cuza” Iași

Institutul de Cercetări pentru Inteligență Artificială
Academia Română, București

Institutul de Informatică Teoretică
Academia Română, Filiala Iași

Muzeul Național al Literaturii Române
Filiala București

Intelligentics, Cluj-Napoca

Editura Universității „Alexandru Ioan Cuza” Iași

COMITETUL DE PROGRAM:

Vasile Apopei, Institutul de Informatică Teoretică, A.R., Iaşi

Verginica Barbu Mititelu, Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Corneliu Burileanu, Facultatea de Electronică, Universitatea Politehnică Bucureşti şi Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Monica Busuioc, Institutul de Lingvistică „Iorgu Iordan - Al. Rosetti", A.R., Bucureşti

Alexandru Ceauşu, Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Lucian Chişu, Muzeul Naţional al Literaturii Române, Bucureşti

Constantin Ciubotaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău

Svetlana Cojocaru, Institutul de Matematică şi Informatică, Academia de Ştiinţe a Moldovei, Chişinău

Alexandra Cornilescu, Facultatea de Litere, Universitatea Bucureşti

Dan Cristea, Facultatea de Informatică, Universitatea „Al. I. Cuza" şi Institutul de Informatică Teoretică, A.R., Iaşi

Nicolae Curteanu, Institutul de Informatică Teoretică, A.R., Iaşi

Gabriela Czibula, Universitatea „Babeş-Bolyai", Cluj-Napoca

Cristina Florescu, Institutul de Filologie Română „Al. Philippide", A.R., Iaşi

Corina Forăscu, Facultatea de Informatică, Universitatea „Al. I. Cuza", Iaşi şi ICIA, A.R., Bucureşti

Maria Georgescul, ISSCO / TIM, ETI, Université de Genève

Gabriela Haja, Institutul de Filologie Română „Al. Philippide", A.R., Iaşi

Catalina Hallett, Open University, London

Sanda Harabagiu, Department of Computer Science University of Texas at Dallas

Maria Husarciuc, Centrul de Studii Biblico-Filologice „Monumenta linguae Dacoromanorum", Universitatea „Al. I. Cuza", Iaşi

Adrian Iftene, Facultatea de Informatica, Universitatea „Al. I. Cuza" Iaşi

Diana Inkpen, Université d'Ottawa

Radu Ion, Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Elena Irimia, Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Doina Jitcă, Institutul de Informatică Teoretică, A.R., Iaşi

Daniel Marcu, University of Southern California

Rodica Marian, Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu", A.R., Cluj-Napoca

Dan Moldovan, University of Texas at Dallas

Vivi Năstase, EML Research, Heidelberg

Constantin Orasan, University of Wolverhampton

Oana Postolache, ISI - University of Southern California, Los Angeles

Irina Prodanoff, ILC-Pisa şi Università di Pavia

Marius Paşca, Google Research

Georgiana Puşcaşu, University of Wolverhampton

Vasile Rus, University of Memphis

Vitalie Scurtu, Intelligentics, Cluj-Napoca

Violeta Seretan, Departament de Linguistique, Universite de Geneve

Mihai Surdeanu, University or Stanford

Valentin Tablan, University of Sheffield

Amalia Todiraşcu, Universite Marc Bloch, Strasbourg

Horia-Nicolai Teodorescu, Institutul de Informatică Teoretică, A.R. şi Universitatea Tehnică, Iaşi

Dan Tufiş, Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

Ioana Vintilă-Rădulescu, Institutul de Lingvistică „Iorgu Iordan - Al. Rosetti", A.R., Bucureşti

Adriana Vlad, Facultatea de Electronică, Universitatea Politehnică Bucureşti şi Institutul de Cercetări în Inteligenţă Artificială, A.R., Bucureşti

COMITETUL DE ORGANIZARE:
Vlad Alexa, FII-UAIC (vlad.alexa@infoiasi.ro)
Lucian Chișu, MNLR (lucianchisu@ gmail.com)
Ioana Corâci, (ioana.coraci@yahoo.com)
Marius Corîci, Intelligentics (marius@ intelligentics.ro)
Dan Cristea, FII-UAIC și IIT-AR (dcristea@ info.uaic.ro)
Florin-Tudor Cristea, FII-UAIC (florin.cristea@ infoiasi.ro)
Lucian Gâdioi, FII-UAIC (lucian.gadioi@info.uaic.ro)
Adrian Iftene, FII-UAIC (adiftene@info.uaic.ro)
Angela Ioniță, ICIA-AR (aionita@racai.ro)
Petru-Adrian Istrimschi, MediaEC-UAIC (petru.istrimschi@infoiasi.ro)
Mihai-Alex Moruz, FII-UAIC și IIT-AR (mmoruz@info.uaic.ro)
Marius Răschip, FII-UAIC (mraschip@info.uaic.ro)
Horia-Nicolai Teodorescu, IIT-AR și Universitatea Tehnică, Iași (hteodor@etc.tuiasi.ro)
Diana-Maria Trandabăț, FII-UAIC și IIT-AR (dtrandabat@ info.uaic.ro)
Dan Tufiș, ICIA-AR (tufis@racai.ro)
Eugenia Țarălungă, MNLR (eugenia12@gmail.com)

Cuprins

Cuvânt înainte 9

Capitol I – *Corpusuri vocale şi prelucrarea vorbirii*

Metodologie pentru constituirea şi analiza unui corpus adnotat de semnale vocale - cazul SROL 13
Horia-Nicolai Teodorescu
Tehnici de identificare a zonelor vocalice în secvenţe rostite în limba română 23
Marius Zbancioc, Horia-Nicolai Teodorescu şi Monica Feraru
Aspecte metodologice de organizare a datelor şi de analiză statistică a vocilor emoţionale 35
Horia-Nicolai Teodorescu, Ioan Păvăloi şi Monica Feraru
Program de editare de contururi intonaţionale în limba română bazat pe ierarhii de forme prosodice funcţionale 45
Vasile Apopei, Doina Jitcă şi Otilia Păduraru
Corpus pentru gnatofonie: protocol, metodologie, adnotare 51
Horia-Nicolai Teodorescu şi Alina Untu
Dispozitiv electronic de preprocesare în regim paralel a spectrului vocal 61
Mircea Hulea şi Alina Untu

Capitol II – *Platforme, dicţionare şi corpusuri pentru prelucrarea textelor*

Resurse lingvistice româneşti în flux continuu 73
Dan Cristea
Comunicarea electronică şi problemele noastre ortografice – fără soluţii? 81
Lucian Chişu
Când migrăm la diacriticele corecte? 89
Bogdan Stăncescu
Construcţia automată de corpusuri multilinguale 103
Tiberiu Boroş, Dan Tufiş şi Alexandru Ceauşu
Parsarea comparativă a dicţionarelor-tezaure româneşti, franceze şi germane 113
Neculai Curteanu, Alex Moruz şi Diana Trandabăţ
Realizarea unui treebank românesc 123
Cenel-Augusto Perez
Monitorizarea presei în cadrul proiectului Neorom 131
Ana-Maria Barbu
Emoţii în cuvinte: elaborarea unei resurse multilingve 141
Victoria Bobicev, Victoria Maxim, Tatiana Prodan, Natalia Burciu şi Victoria Angheluş

Capitol III – *Aplicaţii ale tehnologiilor lingvistice textuale*

Sistem întrebare-răspuns antrenabil pentru limba română 153
Dan Ştefănescu, Radu Ion, Alexandru Ceauşu, Dan Tufiş, Elena Irimia şi Verginica Barbu-Mititelu
Un sistem de traducere automată română-franceză 165
Mirabela Navlea şi Amalia Todiraşcu

Servicii Web interoperabile și multilinguale 175
Radu Ion, Alexandru Ceaușu, Dan Ștefănescu și Dan Tufiș
Tipare hiponimice pentru limba română 185
Verginica Barbu-Mititelu
Mecanismele generative ale morfologiei derivaționale 195
Mircea Petic
Dezvoltarea unui parser de roluri semantice pentru limba română 203
Diana Trandabăț și Dan Cristea
Supravegherea pe Internet: părerea consumatorilor despre anumite produse sau despre anumite evenimente 213
Adrian Iftene, Alina-Elena Mihaila, George-Alexandru Vlad și Geta Stancu
Folosirea verbelor pentru determinarea inferențelor textuale 223
Mihai Alex Moruz

Summaries / Abstracts 233

Index de autori 244

CUVÂNT ÎNAINTE

Acest volum include lucrările acceptate la a şaptea ediţie a Conferinţei Consorţiului de Informatizare pentru Limba Română (ConsILR), dedicată *Resurselor Lingvistice şi Instrumentelor pentru Prelucrarea Limbii Române* şi găzduită de Muzeul Naţional al Literaturii Române din Bucureşti, în perioada 6-7 mai 2010. Organizatorii evenimentului au fost Facultatea de Informatică a Universităţii „Alexandru Ioan Cuza" din Iaşi, Institutul de Cercetări pentru Inteligenţă Artificială al Academiei Române din Bucureşti, Institutul de Informatică Teoretică al Academiei Române – Filiala Iaşi, Muzeul Naţional al Literaturii Române din Bucureşti şi firma Intelligentics.

În contextul unor eforturi internaţionale tot mai susţinute pentru realizarea de resurse şi instrumente lingvistice interoperabile (a se vedea în acest sens proiectele CLARIN, FlaReNet, MetaNet, MetaNet4u, Panacea, MultilingualWeb, American National Corpus, Asian LanguageGrid şi multe altele) conferinţa ConsILR a fost dedicată resurselor lingvistice şi instrumentelor pentru prelucrarea limbii române (scrise şi vorbite). Articolele incluse în acest volum reflectă o parte a eforturilor cercetătorilor români de a răspunde mişcării mondiale de susţinere tehnologică a limbilor naturale pentru utilizarea lor în mediile de comunicare electronică.

La Conferinţă au participat experţi lingvişti, informaticieni, matematicieni şi ingineri din domeniile ingineriei limbajului, lingvisticii teoretice, inteligenţei artificiale, a prelucrării semnalelor şi ingineriei medicale, din Bucureşti, Cluj-Napoca, Iaşi, Chişinău şi Strasbourg. Ca şi la ultimele două ediţii, lucrările conferinţei au putut fi urmărite online, de data aceasta însă direct în Internet, prin grija MEDIAEC, Laboratorul Multimedia al Universităţii Alexandru Ioan Cuza, căruia îi mulţumim şi pe această cale. Mulţumim de asemenea celorlalţi doi sponsori ai manifestării, UPC Business şi România liberă. Programul complet al Conferinţei şi înregistrările audio-video pot fi consultate la adresa http://consilr.info.uaic.ro/consilr2010/.

Un număr de conferinţe invitate au focalizat interesul publicului spre probleme legate de semantică lexicală, utilizarea diacriticelor în limba română, crearea de resurse româneşti de interes naţional (cum sunt colecţia de dicţionare grupate sub numele Dex-online şi secţiunea românească a Wikipedia). Doar o parte din aceste conferinţe au fost transmise editorilor pentru a fi incluse în prezentul volum. Partea finală a Conferinţei s-a axat pe probleme de lexicografie computaţională, legate de realizarea marelui Dicţionar Tezaur al Limbii Române în format electronic, eDTLR.

Volumul cuprinde 22 de lucrări şi este structurat în trei părţi: Capitol I – *Corpusuri vocale şi prelucrarea vorbirii*, Capitolul II – *Platforme, dicţionare şi corpusuri pentru prelucrarea textelor* şi Capitolul III – *Aplicaţii ale tehnologiilor lingvistice textuale*. Lucrările reflectă direcţiile principale de interes actual în institutele de profil ale Academiei Române, în universităţi şi firme de informatică din ţară, dar şi în câteva locuri aflate în afara României. Unele articole sunt metodologice, altele ilustrează direcţii noi de cercetare şi un număr mare de lucrări prezintă progrese în domeniul creării de resurse lingvistice româneşti. Toate lucrările incluse în volum au fost extinse şi corectate de către autori după Conferinţă, pe baza sugestiilor referenţilor, întrebărilor puse în timpul prezentărilor, precum şi a sugestiilor editorilor. O corecţie finală a

întregului volum a fost realizată de către editori. Volumul este completat de rezumatele lucrărilor în limba engleză şi de un index de autori.

Ediţia a 7-a a ConsILR a marcat un progres în seria acestor evenimente, configurând din ce în ce mai clar maturitatea cercetărilor dedicate tehnologiilor lingvistice aplicate limbii române şi a creării de resurse dedicate. Limba română începe să se contureze ca una din limbile semnificative în privinţa resurselor informatice şi a tehnologiilor aplicate ei. Remarcăm totodată apariţia serviciilor Web lingvistice, care schimbă radical concepţia asupra localizării resurselor şi a aplicaţiilor. Se conturează ideea că deţinerea „în casă" a resurselor şi instrumentelor de prelucrare, adică în calculatorul ori serverul propriu, devine lipsită de interes. Comunicaţiile de mare viteză permit acum conectarea la resurse partajate aflate oriunde pe glob. Acest lucru aduce avantajul uniformităţii adnotărilor, aproape obligând la adoptarea de standarde, împiedică multiplicarea de versiuni diferite şi face posibilă combinarea de resurse şi instrumente de prelucrare din locaţii fizice diferite. Este îmbucurător interesul crescând al Comisiei Europene în privinţa finanţării cercetărilor dedicate tehnologiilor lingvistice şi a bibliotecilor digitale. În articolul său care deschide numărul 9-10 al CLARIN Newsletter (martie-iunie 2010, www.clarin.eu/newsletter/), d-l Kimmo Rossi, director adjunct al Directoratului General pentru Societatea Informaţională şi Media al Comisiei Europene dezvăluie abundenţa de apeluri de proiecte în aceste direcţii ce urmează a fi lansate în toamna şi iarna viitoare. Nivelul finanţărilor de proiecte curente ale Uniunii Europene în direcţia Tehnologiilor Limbajului depăşeşte 70 milioane de euro şi va atinge 120 de milioane de euro în 2012. Deci numai veşti bune!

Editorii mulţumesc autorilor, recenzorilor şi colegilor care au contribuit la apariţia prezentului volum.

Iaşi, Bucureşti, august 2010

Editorii

CAPITOLUL 1

CORPUSURI VOCALE ȘI PRELUCRAREA VORBIRII

METODOLOGIE PENTRU CONSTITUIREA ŞI ANALIZA UNUI CORPUS ADNOTAT DE SEMNALE VOCALE – CAZUL *SRoL*

HORIA-NICOLAI TEODORESCU[1,2]

[1]*Academia Română şi*

[2]*Universitatea Tehnica „Gheorghe Asachi"din Iaşi, Iaşi – România*

hteodor@etti.tuiasi.ro

Rezumat

Descriem o metodologie de constituire şi analiză a unui corpus reprezentativ pentru limba română vorbită. Lucrarea are un caracter în esenţă programatic şi metodologic, dar are şi scopul de a atrage atenţia că un mare număr de corpusuri folosite anterior sau în prezent în inferenţa lingvistică sunt deficitare metodologic după cerinţele actuale şi pot conduce la rezultate invalide.

1. Introducere

Complexitatea şi amploarea proceselor limbilor vorbite, procese influenţate de un foarte mare număr de factori, face ca analizele fonetice – fie ele de natură articulator fonetică, acustic-fonetică, perceptiv fonetică, sau fonologică – să fie adesea bazate pe seturi de date a căror reprezentativitate este incertă sau insuficient caracterizată. Faţă de uzanţele din domeniul foneticii şi dialectologiei „clasice", din secolele trecute, standardele ştiinţifice actuale impun câteva condiţii obligatorii: (i) reprezentativitate statistică; (ii) reproductibilitate; (iii) satisfacerea condiţiilor pentru analiza varianţei. Conform acestor condiţii, foarte puţine dintre studiile din secolele trecute, care la momentul realizării au fost studii de mare profunzime – rămân azi valide ştiintific dincolo de nivelul de observaţii semi-empirice. Ele îşi pierd valoarea de stabilire de fapte fonetice, păstrându-şi doar valoare orientativă, estimativă şi istorică, iar aceasta nu doar din motive de limitări tehnologice (de lipsa de înregistrări de calitate), ci mai ales din motive metodologice în (dez)acord cu cerinţele actuale. În această lucrare, pornim de la condiţiile precizate mai sus şi derivăm un set de imperative metodologice pentru constituirea unui corpus vocal adnotat, cerinţe pe care le-am aplicat la constituirea corpusului SRoL[1]. Condiţiile le încadrăm în două categorii: cele care privesc constituirea primară a corpusului şi validarea lui (Secţiunea 2 a lucrării) şi cele care privesc analiza faptelor lingvistice (Secţiunea 3).

2. Metodologie de constituire a corpusului

2.1. Criterii fundamentale

Criteriul reprezentativităţii statistice impune: (1). Pentru un model de limbă vorbită, un număr de vorbitori suficient de mare, suficient de reprezentativ ca proporţii

[1] Corpusul „Sunetele Limbii Române –SRoL" a fost realizat la iniţiativa noastră în perioada 2004-2010 de un grup care include următoarele persoane (în ordine alfabetică): Monica Feraru, Mihaela Hnatiuc, Raluca Ganea, Ramona Luca, I. Păvăloi, Laura Pistol, H.N. Teodorescu, Diana Trandabăţ, Alina Untu, A. Verbuţă, Oana Voroneanu, M. Zbancioc (şi D. Scheianu, Univ. Piteşti), cu cooperarea unui număr de 53 de persoane care au contribuit cu înregistrări.

(procentaje) pentru populaţia care vorbeşte limba respectivă (unde, prin limbă, înţelegem aici fie întreaga limbă, fie un dialect, fie doar un fenomen lingvistic din limba respectivă), conform criteriilor cunoscute în analiza statistică. (2). Pentru fiecare vorbitor, un număr suficient de repetări ale unei pronunţii, pentru stabilirea reprezentativităţii intra-vorbitor şi a eliminării unor eventuale pronunţii accidental deficitare. (3). Un număr de cuvinte reprezentativ pentru limbă, în sensul că setul include o proporţie semnificativă de silabe frecvente ale limbii, astfel încât să fie satisfăcute condiţiile pentru analiza varianţei la nivelul influenţei contextului asupra pronunţiei vocalelor şi consoanelor. (4). Un set de pronunţii care să reflecte o proporţie neneglijabilă a proceselor limbii, printre altele tonalitatea, încărcătura emoţională şi prozodia generală, chiar dacă fenomenul fonetic analizat este punctual, astfel încât să poată fi determinate influenţele acelor procese. (5). O caracterizare în detaliu a vorbitorilor, astfel încât să se respecte condiţia de verificabilitate (reproductibilitate a analizei) şi să se poată ulterior face analize de varianţă (de determinare a influenţelor diverşilor factori asupra limbii vorbite – v. mai jos).

Criteriul reproductibilităţii impune precizarea cu acurateţe şi complet a metodologiei de culegere de date, a parametrilor instrumentarului folosit, a etapelor de prelucrare primară a înregistrărilor, a algoritmilor folosiţi în prelucrare, precum şi precizarea oricăror alte informaţii necesare pentru reproducerea ulterioară a analizei sau constituirii unui corpus echivalent. De asemenea, criteriul reproductibilităţii impune „publicarea" (facerea publică) a tuturor datelor, începând cu înregistrările şi datele subiecţilor şi terminând cu protocoalele utilizate şi instrumentele proprii dezvoltate pentru prelucrarea şi analiza datelor. Un aspect esenţial al reproductibilităţii îl constituie precizarea tipului de vorbire, conform unui număr de criterii precum: voce educată (cultă) / needucată, voce profesională / neprofesională, monolog / convorbire, voce controlată / necontrolată, caracteristici naturale sau simulate (de ex., încărcătura emotivă), contextul socio-profesional al vorbirii (de ex., conform tipologiei de texte şi comunicări precizată – ne-exhaustiv – de (Turculeţ, 2002, p. 76): „monolog, dialog, povestire, interviu, ştiri, expunere, dispută; discuţie în familie, discuţie amicală, discuţie particulară, discuţie publică particulară (în pauze în instituţie, la piaţă, pe stradă etc. – loc?), discuţie oficială, discurs public, pledoarie juridică" etc. (citare prelucrată). Este indubitabil că, până la crearea de instrumente de adnotare automate, cu performanţe mult mai mari decât cele din prezent, realizarea unui corpus care să cuprindă toate aceste tipuri de voci, cu adnotări corespunzătoare, este foarte dificilă. Ca urmare, este important să fie prezentate clar în corpusuri constrângerile de realizare (deci, tipurile de înregistrări, între altele în raport cu clasele de mai sus, cu toate detaliile posibile). Numai în acest fel se va putea realiza, de ex., un studiu ipotetic dar util, precum „O analiză ‚încrucişată' asupra modificării triunghiului vocalelor în silabele accentuate în discursurile publice ştiintifice faţă de pledoariile juridice şi asupra diferenţelor ce apar în funcţie de sexul vorbitorului în variaţiile dintre discurs ştiintific şi pledoarie".

Criteriul completitudinii (de satisfacere a condiţiilor pentru analiza varianţei) impune cunoaşterea în primul rând a tuturor factorilor despre vorbitor – factori familiari – precum limba maternă a mamei, locul unde a copilărit – educaţionali (locul unde a urmat şcoala primară, nivelul maxim de educaţie atins, profesia – care, ultima, influenţează familiaritatea cu vocabularul – gradul de educare / cultură a limbii folosite), factori sociali, medicali (foarte puţine corpusuri dau informaţii asupra

factorilor medicali, care pot influenţa pronunţia, dar şi folosirea mai largă a limbii, de exemplu ambitusul vocal sau expresivitatea emotivă) etc. În al doilea rând, analiza varianţei presupune ca în corpus să se regăsească toţi factorii importanţi de variabilitate a pronunţiei, pentru a se putea face o analiză a cauzelor ce produc o anumită pronunţie[2] şi a distinge între diversele influenţe. Deci, orice corpus trebuie să includă informaţii complete despre trei categorii: (A) despre vorbitori; (B) despre contextul şi tipul vorbirii (monolog, privat, înregistrare de laborator etc.), precum şi data realizării înregistrării; (C) despre (C1) tehnica înregistrărilor, (C2) criteriile de validare şi acceptare a înregistrărilor; (C4) modul de pre-prelucrare a semnalelor, (C5) modul de segmentare şi (C6) adnotare, (C7) modul de extragere a caracteristicilor şi (C8) de validare şi eliminare a valorilor „anormale" ale caracteristicilor; (C9) modul de prelucrare statistică a datelor, inclusiv instrumentele utilizate[3]; (D) Analiza lingvistică a cuvintelor, propoziţiilor, frazelor înregistrate. Pentru fiecare factor implicat, sunt necesare minim 5, preferabil 10 înregistrări pentru fiecare sex (fiecare cu minim trei repetări).

Criteriul eticii ştiintifice impune păstrarea securizată a datelor personale ale vorbitorilor, informarea şi acordul scris al vorbitorilor pentru a participa la teste şi a face publice înregistrările, precum şi validarea de către un for competent a reflectării cerinţelor eticii în producerea şi analiza corpusului.

Din câte ştim, SRoL este până în prezent singurul corpus care satisface toate aceste criterii pentru limba română – şi unul dintre puţinele pe plan internaţional.

2.2. Vorbitorii şi fişa vorbitorului

Pentru fiecare vorbitor, s-a întocmit o fişă a vorbitorului, fişă care conţine toate datele necesare pentru caracterizarea socio-educaţională şi medicală a vorbitorului. Aceste fişe sunt publice, ca şi restul corpusului de voce, pe situl „Sunetele limbii române", (SRoL). Ca urmare, oricare rezultate publicate de grupul nostru asupra limbii române vorbite, pe baza prelucrării materialului din cadrul SRoL, pot fi verificate de oricine şi deci validate sau invalidate de către alte echipe de cercetare, conform cerinţelor metodice actuale – spre deosebire de marea majoritate a lucrărilor curent publicate, pentru care datele sunt neverificabile (nepublice). În plus, SRoL poate fi complementat de alte grupuri de cercetare cu propriile date şi utilizat în conjuncţie cu alte corpusuri de voce pentru cercetări mai ample.

Fişa include date despre vârstă, sex, zonă geografică în care a copilărit şi s-a format ca vorbitor, zona în care a făcut studiile primare, care se ştie că fixează în mare măsură varianta dialectologică a limbii vorbite, date despre studiile universitare şi locul

[2] Subliniem printr-un exemplu relevanţa acestui deziderat. Să presupunem că se doreşte analiza diferenţelor dintre pronunţiile, pentru o vocală dată, în hiatus, ca vocală glisată şi respectiv ca diftong. Dacă numărul de vorbitori din corpus este redus, de exemplu 6, iar ei au un fundal educaţional-familial foarte diferit (dialectal, social), cu implicaţii în diftongizarea vocalei respective, fundal nedocumentat, în plus dacă printre ei unul are afecţiuni neurologice (nedocumentate) care reduc viteza de reacţie (de ex., viteza impulsului electric pe nervi mai mică decât normal), cu implicaţii în producerea glisandoului vocalic, rezultatele statistice vor indica, indiferent de faptul lingvistic, real sau nu, al diftongizării glisantelor în limba dată, o „tendinţă de diftongizare". Într-o asemenea analiză ipotetică, criteriul de satisfacere a condiţiilor pentru analiza varianţei nu este satisfăcut, iar cauzele personale ale tendinţei de diftongizare sunt „văzute" ca tendinţă a limbii.

[3] Atunci când instrumentele nu sunt publice sau comerciale, de exemplu când sunt dezvoltate de autorii corpusului, instrumentele precum şi algoritmii care stau la baza lor, preferabil şi codul sursă trebuie făcute publice, pentru a îndeplini condiţiile de verificabilitate şi reproductibilitate.

absolvirii, eventualele studii post-universitare, obiceiuri de viaţă care pot influenţa vocea (de ex., fumatul), date biometrice şi patologii cunoscute. Detalii au fost prezentate mai pe larg în alte lucrări. Unul singur dintre vorbitori are patologii cunoscute care să îi afecteze sistemul fonator, respirator, auditiv, sau nervos. Mulţi vorbitori au vocea „educată", fie prin profesia didactică, fie prin obişnuinţa unor prezentări publice.

Studiul s-a făcut cu respectarea în totalitate a păstrării privaţiunii subiecţilor; numele subiecţilor este cunoscut doar realizatorilor sitului şi nu este menţionat în fişa publică a vorbitorului, unde identificarea se face printr-un cod numeric. Toţi vorbitorii au fost informaţi asupra obiectivelor generale ale cercetării şi condiţiilor de difuzare a datelor primare (înregistrările vocale); toţi vorbitorii şi-au dat consimţământul scris pentru înregistrări. Cercetarea a fost avizată sub raport etic de Consiliul Facultăţii ETTI, Universitatea Tehnică „Gheorghe Asachi" din Iaşi.

2.3. Statistica regională şi socio-educaţională a vorbitorilor

Setul de 53 de vorbitori include cca. 65% bărbaţi şi cca 35% vorbitori feminini. Grupa de vârstă reprezentată este majoritar 20-35 de ani (peste 50%) şi doar câţiva vorbitori sunt în grupa de vârstă 45-60 ani. Cu rare excepţii, toţi vorbitorii sunt educaţi şi au copilărit în zona de NE a Moldovei, din Bucureşti, în judeţele Iaşi, Vaslui, Bacău, Botoşani, Suceava. Trei vorbitori sunt din Transilvania, unul din zona Argeş, doi din Vâlcea, doi din Muntenia - Bucureşti, iar unul din Maramureş. Ca urmare, SRoL permite comparaţii între vorbitorii (de limba cultă) din NE Moldovei cu cei din alte regiuni ale României. Profilul socio-educaţional al vorbitorilor este, pentru marea lor majoritate, acela al tânărului născut în anii 1980-1990 şi educat până la nivel de facultate, mulţi având un masterat în informatică, lingvistică, sau în domenii conexe.

2.4. Protocolul de înregistrare

Înregistrările au fost efectuate de mai mulţi membri ai echipei SRoL, folosind programul GoldWave™ 5.0, cu următorii parametri: frecvenţă de eşantionare de 22050 Hz; rezoluţie de 16 şi 24 biţi (utilitarul Praat™ prelucrează numai fişierele pe 16 biţi), monofonie. Programul GoldWave este un program comercial, suficient de performant la nivelul anului scrierii lucrării, care asigură o înregistrare de calitate, cu parametrii doriţi, precum şi o prelucrare primară a datelor la nivel acustic.

2.5. Metodologia de validare a fişierelor de semnal vocal şi de segmentare precisă

Toate fişierele au fost analizate manual, prin ascultare, de echipa SRoL. Cu această ocazie, au fost făcute prelucrări primare, anume au fost eliminate secţiunile zgomotoase sau pauzele prea mari; eventual, întreaga înregistrare a fost rejectată dacă nu satisfăcea cerinţele pentru o înregistrare de calitate, de ex. dacă prezintă „limitări" (trunchieri în amplitudine) de semnal, sau dacă avea zgomot intens, uşor perceptibil auditiv.

Două metode de *segmentare* sunt implicit sau explicit utilizate în SRoL. Prima metodă, numită aici *metoda perceptivă,* constă în ascultarea de către o persoană familiarizată atât cu fonetica generală a limbii respective, cât şi cu instrumentul informatic utilizat. Metoda poate fi considerată a fi o tehnică specifică foneticii perceptive, cu limitele ei – în principal percepţia dependentă de contextul lingvistic în care ascultătorul evaluator s-

a dezvoltat (limba natală, educaţia, cultura lingvistică). A doua metodă[4], dezvoltată semi-empiric de autor de-a lungul anilor este subordonată în egală măsură foneticii acustice şi teoriei semnalelor, deşi este aplicată manual şi subiectiv.

Segmentarea fişierelor de voce a fost realizată manual, spre deosebire de alte studii similare (vezi de ex. (Gendrot & Adda-Decker, 2005)), pentru a se asigura o acurateţe şi un grad de certitudine cât mai mare şi o compatibilitate mare între criteriile perceptive şi cele utilizate în abordarea acustic-instrumentalistă. Segmentarea „de precizie" a fost realizată prin ascultare, conform limitelor percepute între foneme, cu teste repetate pentru a determina cât mai precis graniţele detectate auditiv între fonemele adiacente. Atunci când ascultarea nu permitea o precizare suficient de netă în timp a graniţelor, s-a analizat suplimentar forma de undă şi s-a decis plasarea graniţei acolo unde, vizual, apărea clar trecerea de la o formă de undă a unui sunet la forma de undă a următorului. Atunci când a fost necesar, în special pentru pauze şi zone de tranziţie vocală-consoană, o decizie s-a luat în grup pentru validarea segmentării.

2.6. Metodologia de adnotare

Adnotarea a fost realizată manual, simultan cu etapa de segmentare. Nivelurile de adnotare au fost cele corespunzătoare segmentelor de tip propoziţie – cuvânt – silabă – fonem – segment central de fonem. Au fost distinse la acelaşi nivel cu fonemele trei tipuri de pauze: pauze între propoziţii, pauze între cuvinte, pauze în interiorul cuvântului[5]. Frazele au fost adnotate de echipa SRoL folosind utilitarul Praat™, versiunea 5.1.30 (Boersma, 2002).

La nivel de fonem, segmentele au durate larg variabile, cele mai mici de ordinul câtorva milisecunde (ms), pentru pauze scurte şi pentru plozive, respectiv până la ordinul zecilor de ms, pentru vocale şi fricative prelungite, accentuate. În starea emoţională „tristeţe", duratele monoftongilor sunt sensibil mai mari (cu până la 50%). Variabilitatea duratelor, între foneme şi la acelaşi fonem funcţie de stare, de locul în cuvânt (silaba accentuată fiind tipic mai lungă) etc. face ca „populaţia de ferestre" de analiză să aibă dispersii mari, ca număr de ferestre per fonem.

2.7. Descriere generală a conţinutului corpusului

Sumar, corpusul conţine înregistrări de laborator, controlate, cu zgomot redus, de voci în general culte dar nu cultivate, fără voci profesionale din categoriile artişti, reporteri, avocaţi, dar cu câteva voci de persoane din învăţământ, cu vorbitori fără patologie cronică sau acută, cu vârsta dominantă 20-35 de ani. Propoziţiile sunt pronunţate cu ton neutru sau cu ton emotiv, iar emoţiile au fost simulate şi auto-stimulate. Corpusul conţine vocale susţinute, izolat pronunţate, diftongi, triftongi, hiatusuri, consoane, cuvinte izolate şi propoziţii scurte sau de lungime medie. Specific corpusului este

[4] Metoda constă în urmărirea vizuală a formei de undă şi separarea segmentelor vocalice pe baza apariţiei unei deosebiri substanţiale în formele de undă şi în spectrul acelor segmente, respectiv pe menţinerea în cadrul aceluiaşi segment a zonelor în care forma de undă şi spectrul îşi păstrează un grad mare de similitudine cu zonele anterioare. Această metodă, aplicată vizual de către expertul care realizează segmentarea, a fost dezvoltată într-o metodă obiectivă (instrument informatic) (Teodorescu, 2010 b), folosind patternuri (seturi de trăsături acustice) şi distanţe definite pe spaţiul acestor patternuri.

[5] pauzele intra-cuvânt, nu neapărat între silabe, notate $, pauzele inter-cuvinte: blanc; pauzele de scurtă durată, determinate pe semnal, dar care nu se percep, notate %.

includerea pe de o parte a unor structuri gramaticale particulare, precum apoziția și subiectul dublu, iar pe de altă parte a propozițiilor ponunțate cu încărcătură emoțională. Numărul de pronunții de propoziții cu încărcătură emoțională specificată este, pentru o subclasă (descrisă mai jos) de propoziții, aproximativ egal cu numărul de pronunții cu ton neutru (fără încărcătură emoțională). În acest fel, statistica realizată este echilibrată între tonul neutru (unic în cazurile studiilor anterioare asupra vocalelor limbii române) și pronunțiile emoționale. Ca urmare, statistica prezentată are o mai mare rigoare în privința variabilității pronunțiilor posibile și deci în privința modelului complet al limbii vorbite și, în același timp, capătă o dimensiune suplimentară, a încărcăturii emoționale.

Corpusul conține un set de propoziții cu încărcătură semantică imprecis delimitată și lăsată la latitudinea interpretării vorbitorului. De exemplu, una dintre propoziții poate fi interpretată ca interogativă (în scriere marcată prin ?), interogativ-exclamativă (?!), sau ca rămasă în suspensie (...). Alte propoziții suferă interpretări de tipul simplu afirmativ (.), exclamativ (!), sau pot fi intrepetate în suspensie (...). În acest fel, prin alegerea cu grijă a propozițiilor, s-a asigurat nu doar libertate de intrepretare vorbitorului – și implicit un grad ridicat de naturalețe și expresivitate – dar s-a asigurat și o ponderare implicită rezonabilă între diversele prozodii și încărcături emoționale, aspect original și specific corpusului nostru. Modelul de limbă este mai realist astfel, chiar dacă analiza nu include decât un timp total relativ mic de vorbire.

Studiile viitoare vor putea corecta și crește precizia modelului de limbă vorbită, în principal prin analiza unor înregistrări de durată totală mare, echilibrate conform principiilor formulate în (Teodorescu, 2010a), anume: echilibru între pronunții cu tonalități diferite (neutru afirmativ, neutru exclamativ, neutru interogativ, neutru în suspensie) și cu activări (emoționale) diferite, pentru variate emoții. Acele studii viitoare vor trebui să implice zeci de ore de înregistrări pentru a atinge un grad mai mare de semnificație statistică față de studiul actual[6].

2.8. Tipuri de contexte pentru vocale

Pentru a satisface criteriul completitudinii (de satisfacere a condițiilor pentru analiza variației) este necesar ca în corpus să se afle, pentru fiecare proces studiat și pentru limbă în ansamblu, un număr suficient de cazuri care reflectă condiții diferite. De exemplu, printre condițiile care influențează pronunția vocalelor se numără contextul în care apar vocalele (avem în vedere aici, ca exemplu, determinarea unei statistici pentru triunghiul vocalelor în limbă). Este cunoscut că formantul zero (fundamentala) unei vocale este influențat semnificativ de tonalitatea propoziției, de emoția exprimată, de nivelul de accentuare al cuvântului, de contextul cuvântului și de locul unde se află plasată vocala în cuvânt, influențe reflectate de linia prozodică specifică tonalității și imprimată propoziției ca atare, prin urmare și vocalei. În același timp, este cunoscută influența „contextului imediat", constituit de fonemul (monoftongul) anterior sau ulterior vocalei, de existența unui diftong sau de glisare, asupra formanților, în special asupra lui F_1 și F_2. Pentru fiecare vocală, numărul de ocurențe în diversele contexte (în

[6] Până la apariția unor instrumente a căror precizie de segmentare și adnotare automată (și interpretare) să fie echivalentă cu cea a experților umani (uneori constituiți în echipă, ca în cazul unor analize de finețe făcute de noi), asemenea studii sunt greu de realizat. Într-adevăr, activitatea la corpusul SRoL până în prezent este echivalentă cu peste 2-3 (ani × om), iar un studiu comparabil pentru zeci de ore de înregistrări ar presupune zeci de (ani × om) de activitate.

care apare vocala) semnificative statistic pentru limbă trebuie să fie suficient de mare pentru o inferenţă statistică. În prezent, corpusul acoperă cca. 35% din limba română, la nivelul silabic[7] (probabilitatea de regăsire a unei silabe din limba română în corpus este 35%, altfel spus, suma probabilităţilor silabelor din corpus în l.r. este aproape de 0.35).

Un aspect important al contextului vocalelor este structura prozodică în care sunt pronunţate. În acest sens, distingem contexte prozodice din care vocala face parte, de tip DU (creştere F_0), F_0 constant, respectiv UD (descreştere F_0), eventual primul şi ultimul context cu atributul suplimentar „creştere / descreştere rapidă". Reprezentativitatea SRoL sub acest aspect nu a fost încă determinată.

Statisticile efectuate „încrucişat", „în diagonală" (pentru selecţii de contexte în care apar vocalele şi pentru grupuri de vorbitori) se realizează în aplicaţiile dezvoltate de noi şi de colectivul SRoL prin selectarea acelor instanţieri de vocale care corespund criteriului respectiv. Asupra lor revenim în altă lucrare transmisă la CONSILR2010.

2.9. Tipuri de propoziţii, intonaţii şi încărcături emoţionale

Statistica a fost realizată pe un set de propoziţii prefixate, selectate, după cum s-a mai spus, pentru a satisface riguros criteriul de „compatibilitate multi-emoţională" (Teodorescu 2002 a). Anume, criteriul impune compatibilitatea propoziţiei respective cu interpretări emoţionale pentru toate cele trei emoţii selectate drept relevante pentru corpus (bucurie, furie, tristeţe), alături de tonul neutru[8]. Alegerea stărilor emoţionale a fost realizată după teste preliminare (împreună cu dr. M.S. Feraru) asupra unui set de şapte emoţii; selecţia s-a făcut astfel încât separarea între emoţii să fie suficient de netă (ambiguitate redusă la recunoaştere, deci recognoscibilitate rezonabil de uşoară de către ascultători neutri) şi în acelaşi timp, încărcarea emoţională a propoziţiei să fie suficient de facil de realizat pentru vorbitori. Relevanţa pentru corpus a tipurilor de emoţie este determinată de antagonismele bucurie – tristeţe şi bucurie – furie, de caracterul intens al acestor emoţii, precum şi de uşurinţa cu care pot fi exprimate suficient de distinct una de alta. Propoziţiile incluse în această statistică sunt toate de natură afirmativă; faptul că nici una nu include negaţii este probabil limitarea principală a SRoL în prezent.

Gradul de încărcare emoţională pentru fraze a fost determinat cu ajutorul unei aplicaţii *online* realizată de colectiv (Pistol & Teodorescu, 2010), aplicaţie cu ajutorul căreia au fost colectate opiniile asupra recognoscibilităţii emoţiei în pronunţia respectivă de la un număr mare de evaluatori. Conform scării folosite de (Beller, Obin, Rodet, 2008), „gradul de activare" (a emoţiei), pentru pronunţiile de propoziţii din corpusul SRoL este între inactivare şi activare mare, pe o scară cu patru trepte (inactivare, activare mică, medie şi mare). Ambele tipuri de activare a emoţiei, activare extrovertită şi introvertită sunt prezente în pronunţiile de propoziţii din corpus. Toate emoţiile au fost de tip simulat („acted emotions"), dar „trăit", în sensul că li s-a recomandat subiecţilor să se

[7] Probabilităţile silabelor au fost preluate din modelele de limbă determinate pe corpusuri ample în Adriana Vlad et al., Limba română scrisă ca sursă de informaţie, Ed. Paideia, Buc., 2003

[8] Propoziţii şi interpretările posibile (/ / /): Aseară (. / ! / ...), Vine mama (. / ! / ...), Ai venit iar la mine (! / ...), Cine a făcut asta (? / ?! / ... /), Omul meu îl lucră (. / ! / ...), Oricum, îţi poţi câştiga locul dorit (. / ! / ...). Fraze cu subiect dublu sau apoziţie: Vine ea mama!, „A trecut el aşa un răstimp", (M. Sadoveanu), O şti el careva cum să rezolve asta, Mama vine şi ea mai târziu, Mama ştie ea ce face, Chiar ştie el ce face?

gândească la o situaţie reală care ar elicita emoţia respectivă şi deci pronunţia emoţională dorită[9].

3. Metodologia de analiză

3.1. Metodologia de prelucrare şi analiză formantică (acustică)

După analiza preliminară a calităţii înregistrărilor, acestea au fost filtrate, astfel încât să rămână doar spectrul de frecvenţe de interes în analiza formantică, anume între 70 Hz şi 10 kHz. Filtrarea s-a realizat folosind facilitatea utilitarului GoldWave™ de a se preciza de către utilizator banda de filtrare şi forma caracteristicii filtrului.

Pentru prelucrări spectrale (detecţia formanţilor şi a frecvenţei fundamentale), s-a selectat un instrument informatic dintre cele mai puternice la ora actuală şi în acelasi timp gratuit, utilitarul Praat™, conceput şi pus la dispoziţie comunităţii internaţionale de către P. Boersma şi D. Weenink (Boersma, 2002). Analiza minuţioasă a influenţei deplasării ferestrei[10] asupra rezultatelor (Teodorescu, Feraru) a indicat că alegerea este convenabilă şi nu produce erori la variaţii mici ale pasului de deplasare sau ale dimensiunii ferestrei. Utilitarul Praat furnizează valorile frecvenţelor medii ale formanţilor F_0 [Hz], F_1, F_2, F3, F4 [Hz], pentru zonele delimitate la segmentare, precum şi valori instantanee ale formanţilor pentru fiecare „fereastră" de analiză.

3.2. Eliminarea erorilor

Utilitarul Praat™, deşi printre cele mai performante în prezent, produce cca. 10% erori evidente, din numărul total de determinări de valori medii, la valorile primilor doi formanţi şi aproape acelaşi procent de erori la determinările de frecvenţă fundamentală. Deoarece valorile „evident eronate" la care ne referim sunt adesea mult mai mari decât cele credibile, aceste erori afectează semnificativ, chiar la frecvenţe de apariţie de ordinul 5-10%, rezultatele statisticilor. Ca urmare, este necesară eliminarea erorilor grosiere produse de instrumentul de analiză înainte de realizarea statisticii. Dat fiind numărul mare de fişiere, nu se pune problema determinării manuale a valorilor corecte. Soluţia la care am recurs este de a elimina valorile „evident" eronate precum şi pe cele „aberante" (outliers). Statisticile s-au realizat doar pe valorile „valide". Această eliminare a valorilor „anormale" poate conduce la eliminarea, pe lângă valorile eronate şi a unor valori reale, dar anormale, nespecifice populaţiei globale. Aceasta este corect din punctul de vedere al cercetării dacă suntem interesaţi de „procesele medii", de caracterizarea globală şi nu de elemente specifice, rare de pronunţie, care pot să fie mascate de eliminările operate.

În etapa de prelucrare intermediară a datelor, au fost corectate două tipuri de erori, primul produs de instrumentul de analiză Praat™, iar al doilea este posibil a se datora

[9] Gradul de activare al emoţiilor îl determinăm astfel: cuartila superioară (conform evaluării de către ascultatori) este „foarte expresiv", următoarea cuartilă (50-75%) este „expresivă", a treia cuartilă (25-50%) este puţin expre-sivă, iar ultima slab expresivă (ne-expresivă). Facem distincţia între *expresivităţile pentru fiecare emoţie în parte*, deoarece unele persoane pot activa şi exprima uşor o emoţie, de exemplu bucuria, dar nu şi altele, de ex. „triste-ţea". O persoană cu scoruri printre primii 25% dintre vorbitori (scorurile sunt calculate astfel: 4 pct. = f. expresiv, 0 = puţin expresiv; scor personal între 0 şi 16) o considerăm extrovertită, iar între 25%-75%- persoană medie.

[10] Instrumentul Praat a fost setat să lucreze cu ferestre de analiză de 0,025s şi cu suprapuneri ale ferestrei alunecă-toare (suprapuneri între ferestre succesive) de ordinul a 40%, ceea ce corespunde la deplasări de 100 ms ale ferestrei.

instrumentului sau altor factori (înregistrări deficitare, pronunţii deficitare). Aplicaţia informatică dezvoltată permite eliminarea automată a valorilor eronate date de Praat™: (a) nedefinite pentru F_0, acolo unde există sunet vocalic; (b) valori pentru F_0 acolo unde nu este sunet vocalic; (c) valori anormal de mici sau de mari determinate de Praat pentru F_0, F_1, F_2. Anterior efectuării calculelor statistice pentru fiecare formant în parte, din fişierul cu date primare sunt eliminate valorile evident eronate. După această operaţie de curăţare, se determină statistica primară – valoarea medie şi varianţa pentru fiecare formant – şi se elimină valorile „aberante" (outliers), din afara intervalului $\bar{x}-3\sigma, \bar{x}+3\sigma$. În final, pentru valorile rămase, sunt eliminate „cozile" de 3% din populaţie. Statistica formantului fonemului se calculează cu valorile rămase după acest şir de corecţii. Metoda este prezentată pe larg în alte lucrări.

3.3. Metoda de validare a expresiei emoţionale

Validarea expresiei emoţionale a fost realizată în două etape. În prima fază, cinci evaluatori din cadrul colectivului nostru au efectuat o validare a emoţiei reprezentate (exprimate voluntar) în pronunţia respectivă a propoziţiei. Înregistrările care aveau un nivel de expresivitate inacceptabil de scăzut sau o exprimare confuză a emoţiei au fost eliminate în această etapă dintre înregistrări şi înlocuite eventual cu alte înregistrări. În etapa a doua a validării, am recurs la validarea publică, pe Internet. S-a creat o aplicaţie (Pistol & Teodorescu, 2010) cu ajutorul căreia un utilizator oarecare de Internet, vorbitor de limba română, poate accesa şi asculta oricare dintre propoziţii şi preciza dacă acea propoziţie are încărcătură emoţională, dacă emoţia este evidentă (nu produce confuzie, indecizie asupra tipului de emoţie) şi care este emoţia respectivă.

4. Discuţie şi concluzii

Criteriile expuse au fost aplicate la realizarea corpusului SRoL. Acest corpus nu satisface decât parţial criteriul privind completitudinea necesară pentru analiza varianţei, dar, prin aplicarea coerentă a criteriilor până acum şi în continuare la dezvoltarea corpusului, premisele corectitudinii metodologice ale SRoL sunt asigurate. Nesatisfacerea criteriului completitudinii este relativă şi este datorată numărului încă redus de pronunţii pentru unele tipuri de ocurenţe – de exemplu, pentru vocale glisante şi hiatusuri. Dar în cazul SRoL limitarea nu este datorată, ca în cazul altor corpusuri, necunoaşterii (ne-documentării) unor factori care pot influenţa pronunţia – de la factori specifici vorbitorului la modul de înregistrare sau pre-procesare a fişierelor. La aceste din urmă corpusuri, eroarea este fundamentală şi necorijabilă retroactiv.

Încheiem cu un exemplu de aplicaţie deja iniţiată de noi. Să considerăm problema determinării „triunghiului vocalelor" $(F^v_{1(k)}, F^v_{2(k)})$ pentru limba română[11]. La nivelul actual, această determinare presupune: determinarea „norilor" vocalelor, a elipselor de încredere, global pe întreaga populaţie (masculin şi feminin), separat pe sexe, separat pentru vocale izolate, vocale în cuvinte izolate, în propoziţii, influenţa emoţiilor asupra deplasării triunghiului formanţilor (deplasare cunoscută în literatură, pentru alte limbi),

[11] Din nefericire, l. română este una dintre puţinele limbi europene pentru care încă nu există decât studii semi-empirice (Rosetti, Lăzăroiu, 1982), (Teodorescu et al. 1986), invalide la nivelul cerinţelor prezente, privind triunghiul formanţilor. V. şi (Teodorescu, Feraru, Trandabăţ, 2005).

compararea triunghiului pentru vocale susţinute cu cel obţinut pe întreaga limbă (deci, ca medii ale mediilor pe vocale în diversele lor ipostaze), analiza efectului accentului în cuvânt, analiza efectului contextului (CVC, CV_, _VC, V.V, diftongului, glisării vocalei etc.), compararea cu alte limbi din aceeaşi familie, eventual analiza pe dialecte, profesii etc. La acestea se adaugă analiza modificării funcţiilor densitate de probabilitate şi a parametrilor lor (varianţă, asimetrie, aplatizare) pentru cazurile enunţate mai sus. Într-un sens, acesta este şi (parţial) programul echipei SRoL pentru viitorul apropiat.

Mulţumiri. Autorul multumeşte tuturor colegilor care au lucrat la corpusul SRoL pentru nenumărate şi îndelungi discuţii. Activitatea la SRoL a fost parţial sprijinită de către Academia Română.

Referinţe bibliografice

Beller, G., Obin N., & Rodet X. (2008). Articulation Degree as a Prosodic Dimension of Expressive Speech. Speech Prosody 2008, ISCA, Campinas, Brazil, May 6-9, 2008, 681-684.

Boersma, P.P.G. (2002). Praat, a system for doing phonetics by computer. Glot International, Vol. 5 No. 9/10, p. 341-345.

Gendrot, C. & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. INTERSPEECH 2005, Sept. 4-8, Lisbon, Portugal, 2453-2455.

Pistol, L. & Teodorescu, H.N. (2010). A Note on Testing the Recognition of Emotional States and Tones in Speech. Memoirs of the Romanian Academy, 2010 (under press).

Rosetti, Al., Lăzăroiu, A. (1982). Introducere în fonetică. Ed. Ştiinţifică & Encicl., Bucureşti, 1982.

SRoL, Voiced Sounds of the Romanian Language Project, (autori: Teodorescu, H.-N., Trandabăţ, D., Feraru, M., Ganea, R., Verbuţă, A., Zbancioc, M., Hnatiuc, M., Voroneanu, O., Pistol, L., Untu, A., Păvăloi, I.), www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/voci_emoţionale.htm.

Teodorescu, H.-N., Buchholtzer, L., Poşa, C. (1986). Comunicarea orală om-maşină. Cap. 2 – Vocea şi vorbire, şi Cap. 4 – Sinteza vorbirii, Editura Tehnică, Seria „Tehnica la zi", Bucureşti, 1986.

Teodorescu, H.-N., Feraru, M., Trandabăţ, D. (2006). Situl 'Limba Română Vorbită'. Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române. Editori: Corina Forăscu, Dan Tufiş, Dan Cristea Iaşi, 3 nov. 2006, Editura Universităţii Al.I. Cuza Iaşi, 3-7.

Teodorescu, H.-N. (2010 a). Noisy Speech Files Perceptual Mining for Speech and Noise Patterns, PROMISE, 29 martie 2010, Iaşi.

Teodorescu, H.-N. (2010 b). AI Tools for Speech Analysis Applied to the Romanian Language. (Plenary paper), ECC 2010, 20-22 aprilie 2010, Bucureşti.

Turculeţ, A. (2002). Tipuri de texte orale. In: Klaus Bochmann, Vasile Dumbravă (Eds.), Limba română vorbită în Moldova istorică, Volume 1. Leipziger Universitätsverlag, 2002, pp. 53-78.

TEHNICI DE IDENTIFICARE A ZONELOR VOCALICE ÎN SECVENȚE ROSTITE ÎN LIMBA ROMÂNĂ

M. ZBANCIOC [1,2], H.N. TEODORESCU [1,2], M. FERARU [1]

[1]*Institutul de Informatică Teoretică, Academia Română – Filiala Iași, România*
[2] *Universitatea Tehnică „Gheorghe Asachi", Facultatea de Electronică, Telecomunicații și Tehnologia Informației, Iași – România*

{hteodor, zmarius}@etti.tuiasi.ro

Rezumat

Se prezintă un set de tehnici de segmentare utilizate pentru identificarea zonelor vocalice. Segmentarea este folosită ulterior de instrumentele de extragere a frecvenței fundamentale F_0 și a valorilor formanților F_1,...,F_4. Comparăm precizia segmentării instrumentului propus cu cea a utilitarului Pratt, folosind fișiere adnotate cu mare precizie. Pentru reducerea timpului de rulare s-a optimizat calcul funcției de autocorelație, prin aplicarea unor algoritmi recurenți.

1. Introducere

Segmentarea automată a semnalelor vocale, recunoașterea automată a vorbirii, a limbii de proveniență, identificarea vorbitorului sunt domenii de cercetare cu vechime de câteva decenii, dar încă de mare actualitate. Faza de segmentare este importantă deoarece erorile acesteia afectează în mod direct performanțele extractorului de informații prozodice. Deși în literatura de specialitate se găsesc numeroase articole ce descriu diverse tehnici de segmentare automată (Rabiner & Schafer, 1978), (Calliope, 1989), (Rowde, 1991), problema segmentării nu este complet rezolvată, datorită cvasi-periodicității semnalului vocalic și a gradului mare de variabilitate a caracteristicilor fonemelor de la o limbă la alta.

În (Vidal & Marzal, 1990) se face o trecere în revistă asupra tehnicilor de segmentare insistând asupra metodelor fără constrângeri în ceea ce privește variația contururilor spectrale (SVF), metode ce folosesc o segmentare multi-nivel și o descompunere temporară pentru găsirea limitelor segmentelor. Tehnicile de segmentare combinate cu metodele de recunoaștere automată folosesc HMMs (Hidden Markov Models), parametri acustici, cum ar fi coeficienții MFCC (Mel Frequency Cepstral Coefficients), LPC (Linear Predictive Coding), tehnici de aliniere dinamică în timp (DTW engl. Dynamic Time Warping), etc. (Rabiner & Juang, 1993), (Esposito & Aversano, 2005).

(Juneja & Espy-Wilson, 2002) folosesc HMMs combinate cu SVMs (Support Vector Machines) pentru a detecta vocalele, consoanele finale, consoanele fricative și sonante, respectiv zonele de pauză, folosind 39 de parametri extrași din cepstru. (Matousek et al., 2003) folosește HMMs cu coeficienți spectrali MFCCs raportând procentaje foarte bune de 96% în acuratețea segmentării pe un corpus de date lingvistice pentru limba cehă. (Salam et al., 2009) propune o fuziune a două metode de segmentare a vorbirii, anume metode statistice bazate pe un algoritm de divergență și metode conecționiste de învățare adaptivă MLP (Multi-Layer Perceptron). (Sarkar & Sreenivas, 2005) utilizează

o metodă bazată pe nivelul ALCR (Average Level Crossing Rate) pentru a detecta schimbările temporare semnificative în semnal. Sunt utilizate valori adaptive în funcţie de SNR (Signal-to-noise ratio) şi se compară performanţa de segmentare automată cu fişiere segmentate fonetic manual. Colectivul nostru de cercetare a făcut comparaţii între fişierele de sunet sintetizate şi fişierele de voce naturală (Teodorescu et al., 2009), folosind intervalele de timp specificate prin fişiere de adnotare, pentru a observa diferenţele care apar între acestea la nivelul parametrilor extraşi (durate, valori formanţi, variaţie formanţi). Corpus-ul „Sunetele Limbii Române SRoL" conţine vocale susţinute, fraze pronunţate cu diverse stări emoţionale, sunete gnatosonice, adnotări, instrumente de analiză a semnalului vocalic, care sunt disponibile on-line (Teodorescu et al., 2005).

Pentru determinarea frecvenţei fundamentale (F_0) există două tipuri de metode: metode indirecte şi metode directe. Metodele directe de detecţie a lui F_0 sunt: metoda impedanţmetrică (electroglotograma), metoda cinematografică (stroboscopică), metoda extracţiei F_0 pe baza formei de undă. Metodele indirecte implementate de către noi sunt: metode de analiză în domeniul timp (autocorelaţia, AMDF – Average Magnitude Difference Function), metode de analiză spectrală (metoda cepstrală, HPS – Harmonic Product Spectrum). Nici metodele directe nu pot fi considerate metode absolute, deoarece elementele elastice precum corzile vocale prezintă o vibraţie amortizată.

Pentru validarea metodelor indirecte de extracţie a frecvenţei fundamentale este necesară o comparaţie între rezultatele obţinute în cazul aplicării acestora şi cele obţinute în cazul utilizării metodelor directe. În acest scop am realizat adnotări de mare precizie folosind metoda bazată pe forma de undă. Lucrarea prezintă perfecţionări ale instrumentului expus sumar în (Teodorescu et al., 2007), (Zbancioc, 2006). Scopul cercetării este implementarea unor metode de detecţie de F_0 care să furnizeze rezultate mai bune decât utilitarul Praat™ sau alte utilitare existente.

2. Descrierea instrumentului de analiză prozodică

Instrumentul dezvoltat pentru extragerea informaţiei prozodice conţine mai multe blocuri funcţionale corespunzătoare celor trei etape de procesare a semnalului vocal: preprocesarea, extragerea traseului intonaţional pe baza valorilor frecvenţei fundamentale, extragerea valorilor formanţilor superiori (Fig.1).

În etapa de preprocesare se realizează filtrări ale semnalului cu scopul de a elimina zgomotul nedorit şi de limitare a benzii de frecvenţă în care se caută valorile formantice. De asemenea, în această etapă, un algoritm de segmentare permite eliminarea zonelor consonantice şi a celor de pauză între rostiri şi extragerea zonelor vocalice. Doar pentru aceste secvenţe vocalice, care corespund vocalelor limbii române şi consoanelor semivocalice are sens să fie realizată detecţia lui F_0 şi a formanţilor.

Acest instrument de analiză este un sistem hibrid neuro fuzzy prin blocul decizional ce determină F_0, prin ponderarea rezultatelor furnizate de cele patru metode diferite de extragere a frecvenţei fundamentale (metoda autocorelaţiei 55%, metoda cepstrală 35%, metoda diferenţelor AMDF 5% şi metoda HPS 5%). Ponderile utilizate reflectă performanţele fiecărei metode, estimate ca număr de detecţii eronate. Am asociat ponderi mai mici metodelor cu o probabilitate mai mare de a furniza date eronate. Detecţiile eronate (în principal prima subarmonică, respectiv primele armonici ale lui

F_0) sunt găsite prin compararea valorilor de ieşire consecutive furnizate de aceeaşi metodă şi/sau de alte metode, realizându-se şi o corecţie a acestor „false" detecţii de F_0.

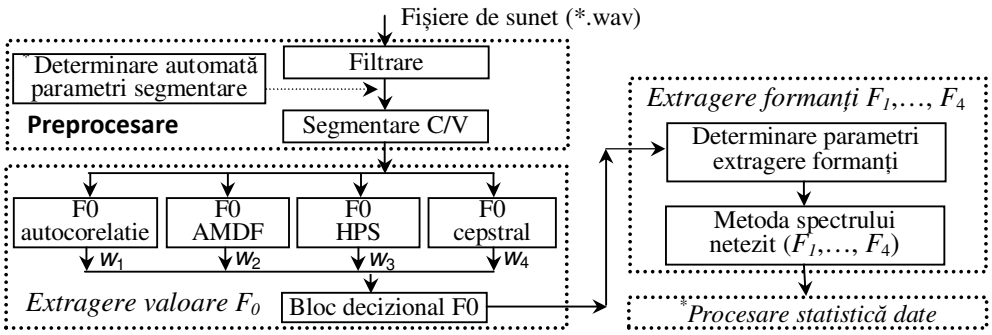


Figura 1: Schema bloc sistem hibrid neuro-fuzzy de extragere informaţii prozodice

Etapa de determinare a valorilor formanţilor superiori se bazează pe acurateţea detecţiei frecvenţei fundamentale prin stabilirea unor intervale de căutare a fiecărui formant în funcţie de F_0. Sunt folosite tehnici fuzzy de concatenare a „spectrelor netezite" rezultate din cepstrele obţinute cu diferiţi parametri, dar şi asocierea de coeficienţi de apartenenţă pentru fiecare candidat găsit în benzile (intervalele fuzzy) de căutare a fiecărui formant.

Ideea utilizării unui bloc decizional hibrid este, din câte ştim, originală şi a permis obţinerea de rezultate mai bune la o inspecţie vizuală a liniei prozodice, fată de alte instrumente similare de analiză a informaţiei prozodice existente pe web: Klatt, Praat (Boersma & Weenink, 2006), Goldwave (www.goldwave.com), Wasp (www.wasp.dk), Speech Analyzer (www.sil.org/computing/sa), Winpitch (www.winpitch.com).

3. Tehnici de segmentare C/V

În această lucrare prezentăm doar prima fază de execuţie a instrumentului de analiză prozodică şi anume etapa de preprocesare şi segmentare, care trebuie să se realizeze cu cât mai puţine erori posibile, deoarece toate etapele ulterioare se bazează pe aceasta. Erorile din faza de segmentare vor afecta în mare măsură eroarea finală de detecţie. Determinăm erorile de segmentare comparând segmentele vocalice identificate automat cu cele marcate manual într-un fişier adnotat de mare precizie, folosind utilitarul Praat. Metodologia de notare şi marcare a fonemelor este descrisă în cele ce urmează.

3.1 Metodologia de adnotare

Adnotarea prin ascultare este subiectivă şi greu de realizat cu precizie, fiind necesară în paralel şi inspectarea vizuală a formei de undă şi a spectrului semnalului (spectrogramei). Pe baza informaţiilor legate de periodicitatea semnalului, de vârfurile spectrale şi tranziţiile acestora, de componenţa în frecvenţe înalte se poate realiza delimitarea fiecărui fonem. Chiar şi utilizând toate aceste informaţii, stabilirea cu exactitate a acestor limite este dificilă, mai ales datorită perioadelor de tranziţie dintre foneme, a perioadelor de amortizare, a zgomotelor introduse de aerul aspirat/expirat, de sunetele produse la închiderea buzelor etc.

Deoarece adnotările uzuale nu au o precizie suficientă (de ordinul ms), a fost necesară realizarea manuală a unei adnotări de mare precizie, prin metoda (practicată frecvent de al doilea autor) a comparării formei de undă în domeniul timp şi domeniul frecvenţelor.

Durata minimă a pauzelor intravorbire sau a secvenţelor care pot fi sesizate de urechea umană este de ordin zecimi milisecunde, şi de aceea pentru a se ajunge la precizia de ordin ms este necesară şi analiza vizuală a semnalului. Scopul autorilor este de a determina erorile date de alte instrumente, în cazul de faţă Praat, şi de a le compara cu cele date de extractorii de F_0 implementaţi de colectivul SRoL. În acest scop s-au adnotat 10 fişiere din baza de date a sit-ului SRoL (fişierele de sunet provin de la 6 vorbitori, 3 de gen feminin şi 3 de gen masculin). Faţă de metodologia de adnotare aplicată uzual, în plus s-a ţinut cont în procesul de adnotare de următoarele:

- S-au introdus notaţii suplimentare care să permită specificarea sunetelor specifice limbii române (‘â’ = ‘a-‘, ‘ă’ = ‘a+’, ‘ş’ = ‘sh’, ‘ţ’ =’ tz’).
- S-au marcat în mod diferit zonele de pauză astfel: pauzele intervorbire (între rostiri de propoziţii) cu ‘ ‘ (caracterul blanc), pauzele intravorbire (silabe, cuvânt) cu ‘$’, pauzele care nu se aud (prezente în consoanele ‘p’, ‘t’, ‘c’ etc.) cu ‘%’.
- S-au stabilit intervalele de demarcaţie ale fiecărui fonem atât prin inspecţia vizuală a formei de undă (pentru a observa periodicitatea semnalelor în cazul în care acestea sunt vocalice), analiza spectrului şi spectrogramei, cât şi prin inspecţie auditivă.
- S-a validat de mai mulţi evaluatori delimitarea fonemelor şi a pauzelor intravorbire.

3.2 Descriere algoritmi de segmentare C/V

Algoritmii utilizaţi anterior în segmentare estimau pentru fiecare fereastră de analiză (de dimensiune uzuală *N*=512 sau 1024 eşantioane) energia în domeniul timp şi energia spectrală a frecvenţelor joase şi comparându-le cu nişte valori de prag stabileau dacă acele segmente erau vocalice sau consonantice. Etapele algoritmului sunt următoarele:

1) Aplicarea unui filtru Butterworth de ordin 11 în banda [70,6000]Hz (se păstrează informaţia din banda de căutare a formanţilor şi se elimină zgomotul indus de reţea).

2) Parcurgerea întregului semnal şi determinarea energiei maxime $E_W = \sum_{i=1}^{N} |s_i|$ a unei ferestre de analiză *W*, cu dimensiune *N* eşantioane. Folosind această valoare $E_{W\max}$ se parcurge din nou semnalul şi se consideră că în zonele în care energia ferestrei curente este mai mică decât 20% din energia maximă nu avem zonă vocalică.

3) Pentru fiecare fereastră de analiză se calculează energia spectrală totală $E_{FFT}{}^{t} = \sum_{f=0}^{Fs/2} |FFT(s_W)|$ şi energia din banda frecvenţelor joase [70,2500]Hz $E_{FFT}{}^{B} = \sum_{f=70}^{2500} |FFT(s_W)|$ (Fs este frecvenţa de eşantionare). Dacă $E_{FFT}{}^{B} < 0.5 \cdot E_{FFT}{}^{t}$, energia corespunzătoare frecvenţelor înalte este mai mare decât 50% din energia spectrală totală, atunci se consideră că zona respectivă nu este vocalică.

Primul criteriu de segmentare a zonelor consonatice/vocalice este o metodă globală al cărui scop este acela de a elimina zonele de pauză dintre pronunţii, unde este prezent doar zgomotul ambiental, precum şi o serie de consoane care conţin zone de pauză (de

exemplu plozivele). Al doilea criteriu de segmentare C/V bazat pe o metodă locală de estimare a ponderii componentelor spectrale are rolul de a elimina consoanele care au o energie a frecvențelor înalte mai mare în raport cu frecvențele joase.

Valorile de prag utilizate în segmentare au fost determinate empiric după mai multe simulări. Algoritmul de segmentare nu funcționează bine, atunci când există zone cu energia în domeniul timp mult mai mare decât în restul pronunției. Astfel de zone apar datorită intonației (accentuarea unui cuvânt sau silabe), sau la exprimarea unei emoții (cum ar fi starea de furie, bucurie). Din acest motiv s-a testat și o variantă de segmentare în care valoarea energetică maximă este calculată pe o zonă locală restrânsă la 0.5 secunde în jurul eșantionului curent.

S-a realizat de asemenea și determinarea statistică a valorilor optime de prag folosind un algoritm de antrenare supervizat, care să determine pe baza seturilor de antrenare reguli de identificare a zonelor vocalice de cele consonantice și de zonele de pauză dintre cuvinte sau fraze. Seturile de antrenare conțineau pentru fiecare fonem: *zcr* (rata trecerilor prin zero pe secundă), *avg_e*, *std_e* - energia medie și deviația standard a energiei în domeniul timp a ferestrelor de analiză, energia spectrală în benzile B1 [70, 500]Hz, B2 [500, 1000]Hz, B3 [1000, 2000]Hz, B4 [2000, 5000]Hz. S-a preferat în locul rețelelor neuronale sau algoritmilor genetici, utilizarea arborilor de decizie, deoarece aceștia furnizează la ieșire un set de reguli, care au în premise valorile prag determinate automat pentru o clasificare cu eroare minimă. Exemplu de regulă obținută cu See5 (www.rulequest.com/see5-win.html) pentru identificarea zonelor vocalice:

```
Rule 1: (270/120, lift 1.4) ;aplicabilă pentru 150 pattern-uri din cele 270
     IF   E_MED > 0.000516
          B4 <= 0.068311
     THEN class 1 (vowel)  [0.555]
```

Folosirea de metode de segmentare cu valori prag nu oferă întotdeauna rezultate bune. De exemplu, aceste valori de prag nu mai pot fi folosite pentru a delimita zonele de pauză de cele în care vorbitorul rostește ceva, atunci când se vorbește încet, sau când persoana a fost amplasată prea departe de microfon (deși înregistrările s-au efectuat după un protocol care prevede o distanță optimă de la buze la microfon), sau când nivelul de zgomot ambiental este prea mare sau fonemul (vocala), de la finalul secvenței rostite are o energie scăzută și/sau o durată de atenuare mai mare. O soluție pentru această problemă este ajustarea pragurilor în funcție de SNR.

Chiar și criteriul de segmentare bazat pe energia frecvențelor înalte nu reușește să determine unele vocale aflate în hiat/diftong, dificultăți fiind întâlnite în acest caz în zonele de tranziție. Un alt caz este cel al fonemului '*a*' aspirat (pronunțat în timp ce aerul este aspirat pe gură) din propoziția '*A trecut el așa un răstimp*' care este încărcat în frecvențe înalte, dar își păstrează traseele formantice uzuale. O serie de foneme consonantice pot fi găsite în unele pronunții ca zone vocalice (de exemplu 'v', 'z', 'r' etc.), dar există situații în care acestea învecinate fiind cu mai multe consoane nu prezintă acea periodicitate a semnalului și sunt clasificate ca și consoane.

În consecință, este nevoie de un algoritm de segmentare care să verifice dacă semnalul este cvasi-periodic. Criteriile de segmentare utilizate anterior ar putea doar furniza informații suplimentare privind clasa de apartenență (vocală / consoană / pauză rostiri) a

semnalului din fereastra de analiză. În aceste condiții s-a preferat utilizarea funcției de autocorelație, care s-a dovedit a fi și cea mai robustă din punct de vedere a erorilor de extragere a lui F_0, într-un algoritm adaptat pentru faza de segmentare:

1) Filtrarea întregului semnal folosind un filtru de mediere de ordin N=31;

$$s_{filt}[k] = \sum_{i=1}^{31} s[k+i]/31 \ , \ k = \overline{1, len\_s - 31},$$

unde len_s reprezintă lungimea semnalului de intrare. Alegerea ordinului filtrului este justificată de rezultatele simulărilor, prezentate în secțiunile următoare.

```
for i = 1 to n_iteratii
```

2) Calcularea vectorului de valori *scorr* prin aplicarea funcției de corelație pentru fereastra curentă de analiză.

3) Căutarea maximului din vector și calcularea valorii $vF0 = Fs / poz_{\max}$. Dacă această valoare este situată între $poz_{start} = Fs/500$ și $poz_{end} = Fs/80$, atunci „*considerăm*" acel segment ca fiind vocalic, altfel $vF0 = 0$.

```
end_for
```

4) Se determină vectorul boolean $\{z(vF0) \mid z : [0,500] \to \{0,1\}\}$ al variabilității semnalului $vF0$, considerând că între două valori consecutive există variabilitate, dacă între acestea avem o tranziție (variație) mai mare $\pm 5\%$.

$$z[k] = \begin{cases} 1 & vF0[k] > 1.05 \cdot vF0[k+1] \ \& \ vF0[k] < 0.95 \cdot vF0[k+1] \\ 0 & altfel \end{cases}, \quad k = \overline{1, len\_s - 1}.$$

5) Dacă în vectorul z avem variabilitate mai mare de $p\%$ atunci considerăm zona respectivă ca nefiind nevocalică. Se obține în final semnalul $\hat{s}_{segm}$

$$\hat{s}_{segm}[k] = \begin{cases} s[k] & , \hat{z}[k] = \sum_{i=-N/2}^{N/2} z[k+i] < N \cdot p\% \\ 0 & altfel \end{cases}.$$

În ultima etapă a algoritmului am considerat pragul de variabilitate de p=5%: dacă din $N = 500$ de valori avem mai puțin de 20 care variază cu $\pm 5\%$ față de valorile vecine, considerăm zona respectivă ca fiind vocalică. Pentru vectorul *vF0* se pot stabili și reguli statistice care să impună ca deviația standard pentru un număr consecutiv de valori să nu depășească o valoare prag dată. Variabila *n_iterații* se deduce în funcție de dimensiunea ferestrei de analiză, W și de pasul de deplasare al ferestrei, $n\_iteratii = \lfloor (len\_s - W)/step \rfloor$. Pentru o segmentare cu o rezoluție maximă a semnalului de ieșire, pasul de deplasare se alege $step = 1$.

În etapa a treia se consideră că o valoare de pe poziția k din vectorul *scorr* este maxim local, dacă este mai mare decât valoarea din dreapta *scorr*[k+1] și din stânga acesteia *scorr*[k-1]. Căutarea maximului corespunzător lui T_0 se face doar în lista maximelor locale astfel găsite. Se evită astfel selectarea ca maxim a primei valori din vector $scorr[0] = \sum_{i=1}^{N} s^2[i]$, care corespunde energiei semnalului din fereastra curentă de analiză.

Un dezavantaj al metodei propuse îl constituie numărul mare de operații care trebuiesc efectuate. Sunt necesare cinci parcurgeri ale unor vectori comparabili ca dimensiune cu semnalul de intrare, s_{filt} obținut după filtrare folosind o fereastră de N=31 eșantioane,

s_{corr} rezultat după calcul funcției de autocorelație cu *N*=1024 eșantioane, z vector variabilitate pentru fiecare două eșantioane consecutive și $\hat{z}$ vectorul variabilității cumulate pentru *N*=500 (folosit la generarea semnalului de ieșire $\hat{s}_{segm}$).

Pentru a obține timpi de rulare mai mici am optimizat algoritmul de segmentare prin folosirea unor funcții recursive de calcul a vectorilor s_{filt}, s_{corr}, $\hat{z}$, esențială fiind scăderea timpului de calcul la nivelul funcției de autocorelație.

3.3 Optimizarea algoritmului de segmentare prin funcții recursive

Funcția de autocorelație necesită un număr mare de operații, timpi mari de calcul, fiind un algoritm de complexitate $O(N^2)$, motiv pentru care s-a evitat anterior utilizarea ei în faza de segmentare pentru a studia periodicitatea semnalelor. Din această cauză, în cazul extractorului de F_0 s-a preferat folosirea unui pas mai mare de deplasare a ferestrei $step = N/16$, unde *N*=1024 reprezintă dimensiunea implicită a ferestrei de analiză.

Pentru rezoluția maximă (*step*=1) s-au obținut pentru un fișier de 10-20 secunde timpi de rulare de ordin 5-10 minute. Prezentăm pseudo-codul algoritmului de calcul a vectorului de autocorelație, pentru fiecare fereastră de analiză, *W,* selectată la parcurgerea semnalul de analizat, *s* cu pas de deplasare al ferestrei $step$.

```
for j=1: step : niter*step
//Calculează C_XX^Wj funcția de autocorelație pentru fereastra curentă W_j
   for k=0:N
      for i=1:P
         C_XX[k] = C_XX[k] + x[i]·x[i+k]
```

În limbajul MATLAB de exemplu, funcția de autocorelație folosește doar valori din fereastra curentă (de dimensiune *N*), motiv pentru care variabila *i* variază între 1: *N-k*. În acest caz formula de calcul pentru funcția de autocorelație devine:

$$C_{XX}[k] = \sum_{i=1}^{N-k} x[i] \cdot x[i+k], \quad k = \overline{0, N}$$

Pentru optimizare am folosit o relație de recurență în calculul funcției de autocorelație a unei ferestre W_2 pornind de la șirul de valori al ferestrei anterioare C_{XX}^{W1}.

$$C_{XX}^{W1}[0] = x_1^2 + x_2^2 + ... + x_N^2 \qquad C_{XX}^{W1}[k] = x_1 \cdot x_{k+1} + x_2 \cdot x_{k+2} + \ldots + x_{N-k} \cdot x_N$$

$$C_{XX}^{W2}[0] = x_2^2 + \ldots + x_N^2 + x_{N+1}^2 \qquad C_{XX}^{W2}[k] = x_2 \cdot x_{k+2} + \ldots + x_{N-k} \cdot x_N + x_{N-k+1} \cdot x_{N+1}$$

$$C_{XX}^{W2}[0] = C_{XX}^{W1}[0] - x_1^2 + x_{N+1}^2 \qquad C_{XX}^{W2}[k] = C_{XX}^{W1}[k] - x_1 \cdot x_{k+1} + x_{N-k+1} \cdot x_{N+1}$$

Relația de recurență de mai sus este calculată pentru cazul în care între cele două ferestre avem un pas de deplasare minimal (*step*=1), fapt ce conduce la o rezoluție maximă a semnalului de prelucrat. Procesul computațional ar necesita în loc de $N \cdot (N+1)/2$ operații de adunare și înmulțire un număr de doar $2 \cdot N$ operații și, teoretic, timpul de calcul ar trebui să scadă de $N/4$ ori. Astfel pentru parcurgerea întregului semnal cu o fereastră de analiză de 1024 eșantioane și calculul funcției de autocorelație am avea teoretic un timp de calcul de aproximativ 250 ori mai mic.

Crescând pasul de deplasare, creşte şi numărul de operaţii de efectuat, astfel încât timpii finali de calcul nu vor fi mai mici decât în cazul în care *step*=1. Aceasta se observă din relaţia de recurenţă rezultată:

$$C_{XX}^{W2}[k] = C_{XX}^{W1}[k] - \sum_{j=1}^{Step} x_j \cdot x_{k+j} + \ldots + \sum_{j=1}^{Step} x_{N-k+2-j} \cdot x_{N+2-j}$$

Pentru filtrarea semnalului s-a preferat folosirea filtrului de mediere neponderat în locul unui filtru digital FIR Butterworth, Bessel, Chebyshev pentru a putea aplica o relaţie de recurenţă. Pentru filtru de mediere ponderat:

$$s_{filt}[k] = \left(\sum_{i=1}^{N} s[k+i] \cdot a[i] \right) \Big/ \sum_{i=1}^{N} a[i],$$

o relaţie de recurenţă devine posibilă doar dacă coeficienţii $a[i]$ sunt egali între ei ($a[0] = \ldots = a[i] = \ldots = a[N]$ obţinându-se un filtru de mediere neponderat):

$$s_{filt}[k+1] = s_{filt}[k] - s[k]/N + s[k+N]/N, \qquad N=31.$$

În aceeaşi manieră s-au folosit funcţii recursive la calculul vectorului de variaţie $\hat{z}$.

$$\hat{z}[k+1] = \hat{z}[k] - z[k] + z[k+N], \quad N=500$$

Folosirea funcţiilor recurente de calcul a condus la obţinerea unor timpi de rulare mai mici de câteva zeci de ori, făcând posibilă aplicarea algoritmului de segmentare propus.

4. Simulări şi rezultate

În urma implementării practice a metodei de segmentare propuse, pentru obţinerea unor erori minime în segmentare s-au formulat următoarele întrebări:
- unde este util să se facă filtrarea: înainte sau după calculul vectorului de autocorelaţie?
- care este ordinul optim pentru filtrul de mediere?
- care este valoarea de prag p% care trebuie folosită pentru obţinerea lui $\hat{s}_{segm}$?

Pentru a răspunde la primele două întrebări s-au folosit semnale armonice de test

$$s = A_1 \cdot \sin(2\pi \cdot f_0 \cdot t) + A_1 \cdot \sin(2\pi \cdot f_1 \cdot t) + A_3 \cdot \sin(2\pi \cdot f_2 \cdot t)$$

în care valorile frecvenţelor f_0, f_1, f_2 se aleg apropiate de valorile frecvenţei fundamentale F_0 şi respectiv ale frecvenţelor primilor doi formanţi, F_1 şi F_2. În aceste condiţii, valoarea de maxim care trebuie detectată şi salvată în vectorul *vF0* ar trebui să fie cât mai apropiată de f_0.

S-au testat două variante de filtrări, una în care se realizează filtrarea înainte (independent) de vectorul funcţiei de autocorelaţie (notată în figurile 2 şi 3 cu (*v2*)) şi una în care se realizează filtrarea în final după calcul vectorului *scorr* (notată cu (*v1*)). Ne interesează acest studiu, deoarece dacă în urma simulărilor s-ar obţine erori de detecţie a frecvenţei fundamentale semnificativ mai mici pentru (*v1*), decât pentru (*v2*) atunci nu s-ar mai putea optimiza algoritmii de segmentare folosind relaţii recurente. În figura 3 este reprezentată eroarea medie de detecţie a lui F_0, pentru mai multe studii de caz $\{F_0,F_1,F_2\}$ = {100,350,900}, {200,550,1100}, {70,350,900}, {100,400,800}, etc., variind ordinul filtrului în mulţimea {1,5,11,15,21,25,31,35,41,45,51}.

Din figura 2 se observă că eroarea de detecție a frecvenței f_0 este minimă pentru un filtru de mediere de ordin 31 și că algoritmul de segmentare funcționează mai bine atunci când funcția de autocorelație este calculată pentru o fereastră de 1024 eșantioane, în varianta *v*2, după ce în prealabil s-a realizat și filtrarea semnalului.

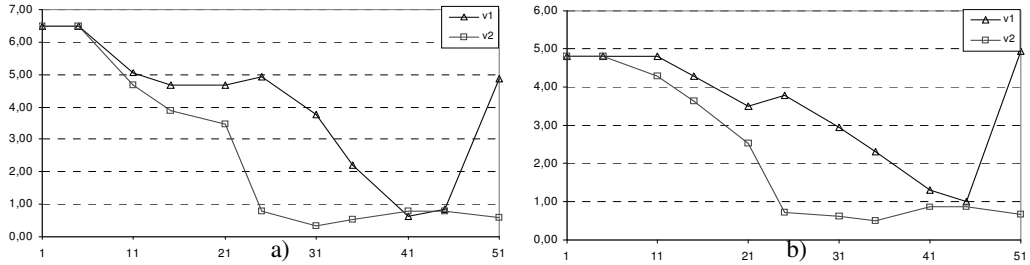


Figura 2: Eroare de detecție a lui F_0 în condițiile în care (v1) filtrarea se face după calcul funcției de autocorelație, sau (v2) filtrarea se face înainte de calculul vectorului *scorr*
a) fereastră de analiză de 512 eșantioane, respectiv b) 1024 eșantioane

Funcționarea optimă a metodei de segmentare cu un filtru de mediere de ordin *N*=31, poate fi explicată prin comportamentul de filtru trece jos FTJ, cu valoarea riplului de aproximativ 520Hz, în condițiile în care banda de lucru este [70-500]Hz.

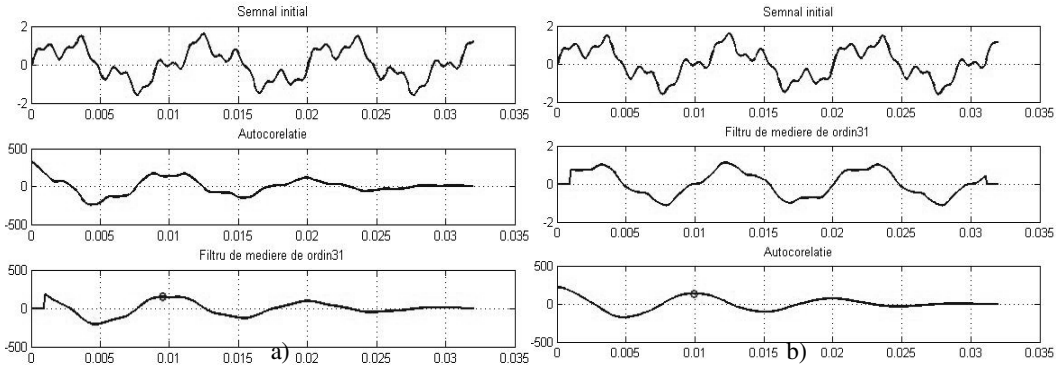


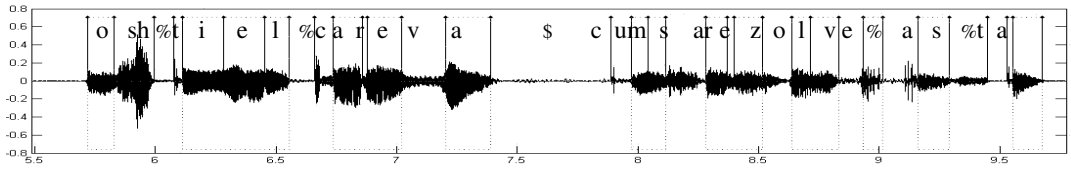
Figura 3: Semnal de test (s=sin(2*pi*t*100)+0.5*sin(2*pi*t*350)+0.2*sin(2*pi*t*900))
a) varianta v1 filtrarea se face după calculul funcției de autocorelație, f0=105.3
b) varianta v2 filtrarea se face înainte, f0=100.6

Un extractor de frecvență fundamentală furnizează următoarele erori: (i) E1: determină F_0 acolo unde nu este sunet vocalic (fals pozitiv); (ii) E2: nu determină F_0 acolo unde este sunet vocalic (fals negativ); (iii) E3: determină eronat F_0 ca valoare. (iv) E4: erori de decizie asupra valorii finale a lui F_0, atunci când avem mai multe metode de detecție.

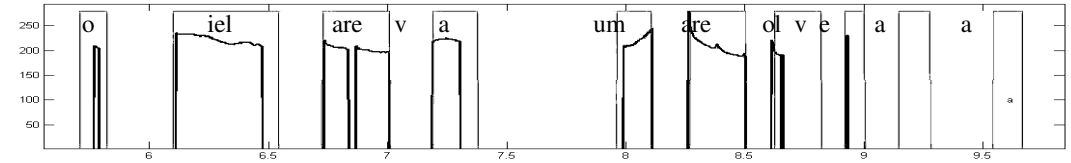
Erorile de segmentare automată sunt specificate de E1 și de E2. La acestea pot contribui și erorile de adnotare (segmentare manuală) în cazul în care aceasta nu este realizată cu precizie. Erorile extractorilor de F0 sunt date de valoarea lui E3, iar la eroarea finală se poate adăuga și eroarea blocului decizional a sistemului hibrid neuro-fuzzy E4.
Semnificative ca valori sunt primele două surse de erori, de care este responsabil blocul de segmentare, E3 fiind eliminate aproape în întregime de algoritmii de corecție, iar valorile lui E4 sunt aproape neglijabile și cuantizabile doar prin inspecție vizuală.
În figurile 3 și 4 sunt prezentate rezultatele segmentării/detecției de F0 cu instrumentele proprii și cu Praat™ pentru aceeași pronunție „*O ști el careva cum să rezolve asta*”) din fișierul de sunet *accent_cuv_urm_v2.wav*. Dacă algoritmii anteriori de segmentare nu

reuşeau să separe toate zonele vocalice, având dificultăţi cu acele foneme care aveau valoarea amplitudinii/energiei mai mică, cu algoritmul nou propus se segmentează foarte bine, chiar şi zone pe care instrumentul Praat nu le detectează. Astfel fonemele '*l*' şi '*e*' din cuvântul '*rezolve'*, vocala '*a*' finală au traseul intonaţional detectat fără discontinuităţi. Nu avem nici false detecţii în cazul fonemelor '*ş*' şi '*c*'. Algoritmul nou propus are câteva zone înguste de ordin ms, situate în zonele de pauză, în care găseşte izolat valori periodice. Pentru a le elimina se pot adăuga restricţii, ca zonele considerate ca fiind foneme vocalice să aibă o durată de minim 5ms, sau restricţii privitoare la energia semnalului care în zonele de pauză sunt semnificativ mai mici decât în segmentele rostite, dacă înregistrarea nu a fost realizată într-un mediu zgomotos.

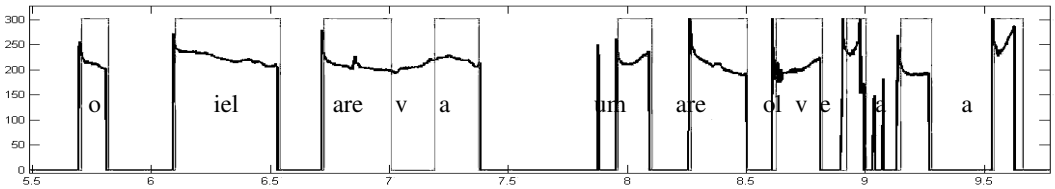
Erorile de segmentare prezentate în tabelul 1 au fost realizate considerând ca zone vocalice segmentele extrase din fişierele de adnotare corespunzătoare vocalelor, diftongilor, consoanelor sonante '*l*','*m*','*n*','*r*'. Nu sunt luate în calcul unele foneme care uneori se comportă atât vocalic (forma lor de undă este periodică), cât şi consonantic. Este cazul fonemului '*v*' care, în '*careva'* este vocalic, dar în '*rezolve'* este consonantic. Zonele de tranziţie dintre foneme vocalice, uneori sunt şi ele nevocalice (de exemplu pentru pronumele '*el*' pronunţat ca regionalism '*iel*', s-au găsit astfel de tranziţii între cele două vocale). Din zonele detectate de program ca fiind vocalice s-au eliminat şi la stanga şi la dreapta o jumătate din fereastră de analiză (N/2 eşantioane), corespunzătoare în general zonelor de tranziţie dintre consoane şi vocale.



a) semnalul iniţial de intrare (pronunţia „O şti el careva cum să rezolve asta")



b) semnalul segmentat cu algoritmi bazaţi pe valori de prag fixe (p1=20% din energia maximă în domeniul timp, p2=50% din energia spectrală)



c) semnal segmentat cu noul algoritm de segmentare (bazat pe studiul periodicităţii semnalului)

Figura 4. Rezultate segmentare şi detecţie F0 (instrumente proprii)

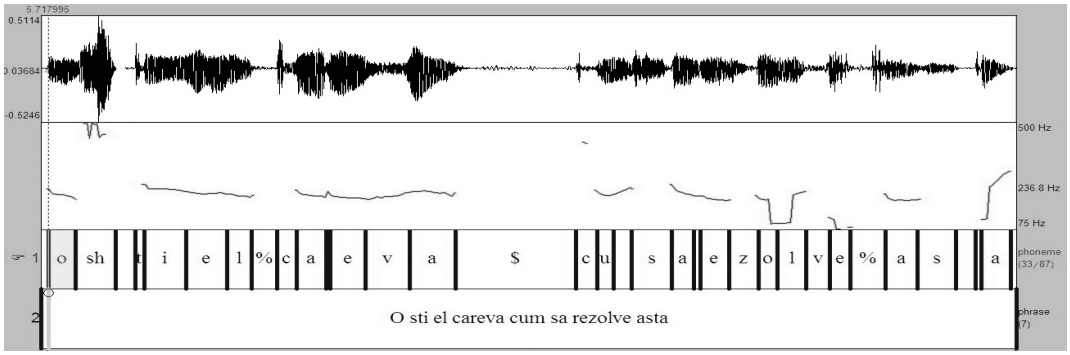


Figura 5. Rezultate segmentare și detecție F0 (soft Praat™)

Tabel 1: Erori extractor de frecvență fundamentală

Fișier de intrare	**F0 - Metoda autocorelației**			**F0 – Praat™**		
	E1	**E2**	**E3**	**E1**	**E2**	**E3**
accent_cuv_urm_v1.wav	0.133	0.064	0.032	0.192	0.011	0.057
accent_cuv_urm_v2.wav	0.084	0.096	0.058	0.153	0.044	0.049
accent_cuv_urm_v12.wav	0.128	0.100	0.065	0.184	0.018	0.763

5. Concluzii. Direcții viitoare

Algoritmul de segmentare propus este mai puțin influențat de amplitudinea și fluctuațiile în amplitudine ale semnalului analizat și prin urmare poate fi adaptat și pentru înregistrări cu un nivel de zgomot mai mare. Este necesară introducerea unor criterii noi de segmentare pentru detecțiile izolate de F_0 din zonele de pauză, respectiv definirea unor valori de prag flexibile, ajustate automat în funcție de SNR (va trebui estimată amplitudinea zgomotului și cea a semnalului util). Se va încerca utilizarea altor parametri care să fie mai puțin influențați de nivelul energetic, cum ar fi rata trecerilor prin zero, pentru semnalul din care s-au eliminat întâi componentele de joasă frecvență.

Conform simulărilor, algoritmii proprii oferă erori mai puține de tipul E1 (F_0 în zone nemarcate ca sunet vocalic) și de tipul E3 (se determină eronat F_0 ca valoare – cu fluctuații). În lipsa unor metode de corecție, soft-ul Praat™ produce „falsele” detecții, ceea ce reprezintă un inconvenient în eroarea globală a sistemului, în comparație cu instrumentele proprii care elimină în proporție foarte mare aceste erori.

Mulțumiri. Cercetarea a fost realizată cu sprijinul Academiei Române, în cadrul temei interne a Institutului de Informatică Teoretică din Iași. Autorii mulțumesc celorlalți co-autori ai sitului Sunetele Limbii Române, precum și referenților anonimi pentru sprijinul și observațiile pertinente.

Contribuția autorilor: Primul autor a implementat instrumentelor de analiză în mediile de programare MATLAB și C++; a elaborat metoda de extragere formanți și optimizarea metodelor de segmentare și extragere a F_0 și a formanților. Al doilea autor a inițiat tema, a coordonat activitatea de cercetare, a elaborat conceptul general de sistem de decizie și metode de extragere F_0 conform cu fig.1, a precizat metoda de adnotare și segmentare manuală cu elemente de noutate privind metoda proprie de combinare a informației temporare cu cea spectrală. Al treilea autor a realizat înregistrările, a efectuat manual adnotările de mare precizie și a identificat problemele de segmentare. Toți autorii au contribuit la analiza rezultatelor și identificarea soluțiilor de îmbunătățire și validarea rezultatelor acestora.

Referințe bibliografice

Boersma, P., Weenink, D., Institute of Phonetic Science, University of Amsterdam, Praat: doing phonetics by computer, *www.praat.org*.

Calliope (1989). *La parole et son traitement automatique*, ISBN 2-225-81516-X, Masson, France.

Esposito, A. & Aversano, G. (2005). Text independent Methods for speech Segmentation, *Lecture Notes in Computer Science*, ISBN 978-3-540-27441-4, 3445, 261-290 (http://www.springerlink.com/content/81fpb3brpq367j7g/).

Juneja, A. & Espy-Wilson, C. (2002). Segmentation of Continuous speech using acoustic-phonetic parameters and statistical learning, *Proceedings International Conference on Neural Information Processing*, (http://www.ece.umd.edu/~juneja/paper1910.PDF), Universitatea din Maryland, Singapore, SUA.

Matousek, J., Tihelka, D., Psutka, J. (2003). Automatic Segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction, *Processing of Eurospeech* 2003, Geneva, 301-304.

Rabiner, L.R., Juang B.H. (1993). *Fundamentals of Speech Recognition* Englewood Cliffs, N.J.

Rabiner, L.R. Schafer R. W. (1978). *Digital Processing of Speech Signal*, Prentice-Hall, Inc. Englewood Clifford, 11-65.

Rowden, C. (1991). *Speech Processing*, McGraw - Hill Book Company, Chapter 2, 35-74.

Salam, M.S., Mohamad, D., Salleh, S.H. (2009). Improved Statistical Speech Segmentation Using Connectionist Appoach, *J. of Computer Science*, ISSN 1549-3636, 5 (4): 275-282.

Sarkar, A. & Sreenivas, T.V. (2005). Automatic speech segmentation using average level crossing rate information, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing: ICASSP*, Philadelphia, SUA, 1, 397-400.

Teodorescu, H.N., Feraru, M., Pistol L., Zbancioc, M. şi alții. (2005). SRoL – Proiectul Sunetele Limbii Române (Voiced Sounds of Romanian Language Project), 2005. [http://iit.iit.tuiasi.ro/romanain_spoken_language/ index.htm]

Teodorescu, H.N., Feraru M., Zbancioc M.D. (2009) Assesing the Quality of Voice Synthesizers, In Burileanu C., Teodorescu H.N. (Eds.), *Advances in Spoken Language Technology*, The Publishing House of the Romanian Academy, Bucharest, România, ISBN 978-973-27-1808-7, 53-66.

Teodorescu, H.N., Trandabat, D., Feraru, M., Zbancioc, M., Luca, R. (2007). A Corpus of the Sounds in the Romanian Spoken Language for Language-Related Education, In Carlos Perinan Pasqual (Eds.), *Revisiting Language Learning Resources*, Cambridge Scholars Publishing (CSP), UK, ISBN 1-84718-156-2, 6, 73-89.

Vidal, E. & Marzal, A. (1990). A review and new approaches for automatic segmentation of speech signals, Signal Processing V: *Theories and Applications*, Torres, L., Masgrau, E., Lagunas, M.A. (eds.), Elsevier Science Publishers B.V. – Universitatea Politehnica din Valencia, Spania.

Zbancioc, M. (2006). Tools for the Archive of the Romanian Language Sounds Project, 4th *European Conf. on Intelligent Systems and Technologies*, ECIT'2006, Iași, Romania.

ASPECTE METODOLOGICE DE ORGANIZARE A DATELOR ŞI DE ANALIZĂ STATISTICĂ A VOCILOR EMOŢIONALE

HORIA-NICOLAI TEODORESCU[1, 2], IOAN PĂVĂLOI[1], MONICA FERARU[1]

[1]*Institutul de Informatică Teoretică al Academiei Române,*

Filiala Iaşi a Academiei Române

[2]*Universitatea Tehnică Gheorghe Asachi din Iaşi*

{hteodor}@etc.tuiasi.ro

Rezumat

Prezentăm o metodologie şi un program pentru analiza statistică a vocilor emoţionale. Prezentăm în principal unele rezultate preliminare privind modul de organizare a datelor în cadrul unei aplicaţii de analiză a adnotărilor pentru fişierele de semnal vocal. Programul permite analiza statistică a caracteristicilor formantice ale vocalelor şi semivocalelor pe subclase de fişiere selectate conform unor caracteristici date de utilizator. Aplicaţia permite un grad de rafinare a analizei emotivităţii în voce la nivelul sunetelor vocalice.

1. Introducere

Analiza expresivităţii emotive (Teodorescu & Feraru, 2007), (Teodorescu & Feraru, 2009) stabilirea unor metodologii teoretice şi a unor algoritmi pentru identificarea stărilor emoţionale (Ververidis & Kotropoulos, 2006), (Nakatsu et al., 1999), (Mcgilloway et al., 2000) au numeroase aplicaţii practice (Scherer, 2000), (Kienast & Sendlmeier, 2008) şi ştiinţifice în domeniul informaticii, lingvisticii, psihologiei etc. Aceste analize sunt însă laborioase şi necesită resurse considerabile, atunci când privesc corpusuri mari de înregistrări de vorbire. În această lucrare prezentăm realizarea unui program care efectuează o statistică generală a formanţilor, cu aplicaţie la unele vocalele („a", „e", „i", „u", „ă") din limba română cu selecţia fişierelor de voce în funcţie de starea emoţională.

Scopul cercetării raportate a fost stabilirea metodologiei de organizare şi de procesare a datelor în vederea prelucrării statistice. Aceasta permite flexibilitate în accesarea informaţiilor, validarea datelor existente, rapiditate în prelucrarea datelor, precum şi dezvoltarea de noi aplicaţii în domeniul procesării limbajului vorbit. Tratarea unitară a volumului mare de date ne permite o abordare eficientă în vederea realizării unor prelucrări statistice complexe. Lucrarea este o continuare a cercetărilor autorilor privitor la analiza şi caracterizarea emotivităţii în voce. Lucrarea este structurată după cum urmează: secţiunea a doua şi a treia sunt dedicate metodologiei de organizare şi de procesare a datelor în aplicaţia dezvoltată de noi (şi numită WinCollection); în secţiunea a patra sunt prezentate rezultatele preliminare obţinute, iar în ultima secţiune sunt descrise concluziile şi direcţiile viitoare.

2. *Metodologia de organizare a datelor*

2.1 Datele primare

Propozițiile înregistrate sunt: „Vine mama”, „Cine a făcut asta”, „Ai venit iar la mine” și „Aseară”. Stările emoționale supuse analizei sunt: bucurie, tristețe, furie și tonul neutru. Exemplificăm analiza pentru un număr de cinci vorbitori (2 feminini și 3 masculini) pentru care avem 63 de înregistrări cu o medie de trei rostiri pe fiecare înregistrare. Există trei înregistrări masculine (bucurie, tristețe și neutru) care au un grad de exprimare a emoției redus comparativ cu ceilalți vorbitori. Toate înregistrările fac parte din corpusul adnotat și documentat SRoL (Teodorescu et al., 2005) și au fost realizate cu o frecvență de eșantionare de 22kHz, cu o rezoluție de 16 biți.

Aplicația este reprezentată de un program realizat în Visual C++. Datele de intrare sunt conținute în patru fișiere: un fișier .wav (înregistrarea propriu-zisă) și trei fișiere .txt (un fișier de adnotare TextGrid și două fișiere cu valori ale frecvenței fundamentale, F0 și ale formanților F1-F3). Adnotarea fișierelor s-a realizat cu utilitarul Praat (Boersma & Weenink, 2006), la nivel de propoziție, cuvânt, silabă și fonem. În procesul adnotării am definit următoarele tipuri de pauze: i) pauzele intravorbire (în rostirea de silabă, cuvânt), care au fost marcate cu simbolul „$”; ii) pauzele intervorbire (între rostiri de propoziții) care au fost codate prin blanc; iii) pauzele care nu sunt percepute și care sunt detectate la nivelul unei analize mai fine; acestea au fost notate cu simbolul „%”.

Atât în fișierele Praat cât și în fișierele de intrare ale programului am codat: â cu a-, ă cu a+, ș cu sh și ț cu tz.

Fișierele text cu valori instantanee ale formanților au fost obținute cu ferestre alunecătoare a căror dimensiune este de 0.025s, iar pasul ferestrei este de 0.01s. Fișierele rulate de program au o durată de maxim 10s, ceea ce corespunde în medie la trei rostiri a propoziției înregistrate. Programul a rulat doar peste fișierele date de utilitarul Praat, dar poate prelucra și alte tipuri de fișiere.

În figura 1, exemplificăm rezultatul rostirii pentru vorbitorul 11861, pentru care avem testate 3 stări emoționale: bucurie, furie și ton neutru. Frazele ale căror pronunții sunt supuse analizei sunt: „Ai v**e**n**i**t **i**a**r** l**a** m**i**n**e**”, „Cin**e** **a** f**ă**c**u**t **a**st**a**”, „V**i**n**e** m**a**m**a**” și „**A**sear**ă**”, iar fonemele analizate sunt „a”, „e”, „i”, „e”, „ă”.

```
Speaker : 11861

        Emotion : bucurie
                Phrase : ai venit iar la mine
                         3 rostiri
                Phrase : aseara
                         3 rostiri
                Phrase : cine a facut asta
                         3 rostiri
                Phrase : vine mama
                         3 rostiri
        Emotion : furie
                Phrase : ai venit iar la mine
                         3 rostiri
                Phrase : aseara
                         3 rostiri
                Phrase : cine a facut asta
                         3 rostiri
                Phrase : vine mama
                         3 rostiri
```

Figura 1: Screenshot referitor la sinteza analizei pentru vorbitorul 11861

2.2 Preprocesare de fişiere

Valorile lui F0 şi F1-F4 au fost calculate automat folosind utilitarul Praat şi salvate în fişierele corespunzătoare vorbitorului, emoţiei şi propoziţiei respective. Există unele situaţii în care F0 apare ca fiind nedefinit. Aceste situaţii sunt următoarele:

- între rostiri - nu există F0 şi nici F1-F4. În tabelul 1 exemplificăm un segment de valori ale lui F0 care este nedefinit şi altul definit pentru propoziţia „Ai venit iar la mine", diftong „ai", prima rostire. Primele valori nedefinite sunt cele care corespund valorilor inexistente din intervalul dinaintea rostirii. Programul ignoră la prelucrare zonele în care valorile sunt nedefinite de utilitarul Praat.
- zone unde nu este sunet vocalic;
- zone unde este sunet vocalic, dar utilitarul Praat nu detectează F0 întotdeauna – eroare de detecţie;
- în cazul în care pe un sunet vocalic se obţin cu utilitarul Praat valori aberante – mult mai mică sau mai mare decât o valoare în limite normale (de exemplu. 70-500Hz, la bărbaţi).

Validarea se realizează prin parcurgerea tuturor datelor de intrare din fişierele de adnotare TextGrid create de Praat, iar pe baza unor criterii, sunt eliminate valori eronate furnizate de acelaşi utilitar. Fiecărui fişier TextGrid (fişier de adnotare asociat unei înregistrări) îi corespunde un fişier cu valorile formantului F0 precum şi un fişier cu valorile formanţilor F1-F4. Corelarea fişierelor se face în mod automat, absenţa unuia dintre ele fiind semnalată ca eroare. Aceste criterii sunt stabilite de către utilizator şi ele pot fi modificate luând în considerare intervalele de variaţie ale fundamentalei şi a formanţilor. Segmentele existente în fişierul de adnotare cu markerii respectivi de timp permit validarea sau invalidarea existentei formanţilor. De exemplu, dacă o pauză este corespunzătoare unui segment între t1 şi t2 (markeri de timp), în fişierul de formanţi pe acelaşi segment se invalidează F0 şi formanţii F1-F3.

Adăugarea de noi informaţii se realizează prin reprocesarea datelor. Fişierul binar conţine informaţii despre data creării, versiunea programului WinCollection utilizat, versiunea fişierului binar, despre persoana care a creat colecţia, precum şi toate informaţiile primare. Un alt avantaj al organizării datelor îl constituie posibilitatea reproductibilităţii rezultatelor obţinute.

Cumularea informaţiilor într-un singur fişier asigură flexibilitate în accesarea lor, selectarea rapidă a lor pe baza unor criterii de intrare stabilite de utilizator. Nu există date primare, chiar şi valorile eronate furnizate de Praat, care să nu se regăsească în colecţie.

Structurarea ne permite analize statistice complexe pe un set de date provenit dintr-o selecţie a lor şi obţinerea rapidă a informaţiilor asupra variaţiei parametrilor analizaţi. În această lucrare raportăm numai statistici pentru vocale. Sunetele semivocalice (despre care se ştie că au F0 şi formanţi superiori) nu sunt luate în considerare în rulările pentru această lucrare, rulări în care s-au setat ca zone de interes doar vocalele.

3. Metodologia de procesare a datelor

Programul permite realizarea unei „statistici tomografice", în sensul că permite determinarea unor „secţiuni de interes" în statisticile realizate. Programul asigură flexibilitate pentru o analiză statistică pe orizontală cât şi pe verticală. De exemplu, programul permite analiza unui singur parametru (F0) pentru toţi vorbitorii şi pentru toate stările emoţionale, sau analiza rezultatelor obţinute de la un vorbitor în comparaţie cu toţi ceilalţi vorbitori. Analiza este organizată la nivelul stărilor emoţionale, la nivelul vocalelor şi la nivelul valorilor formanţilor din cadrul vocalelor.

Programul WinCollection are ca date de intrare fişierele .wav, fişierele de adnotare TextGrid, fişierele cu valorile formanţilor F0 şi F1-F3, fişierele Codes.txt, Phrases.txt şi Emotions.txt. Fişierul Codes.txt conţine date despre toţi vorbitorii ale căror înregistrări trebuie să fie în directorul respectiv. Fişierul Phrases.txt este necesar pentru a stoca frazele permise în fişierele de adnotare. Fişierul e utilizat în depistarea erorilor de adnotare. Prin compararea stringurilor din fişierul Phrases.txt cu stringurile corespunzătoare adnotărilor, sunt depistate lipsa de litere, inversiuni de litere sau alte erori de acest tip. F0 din figura 2 notează toate fişierele generate de Praat conţinând valorile lui F0 pentru fişierele .wav din directorul respectiv. Similar, fişierele F1-F3 din figura 2 desemnează ansamblul fişierelor generate de Praat conţinând aceşti formanţi. Fişierul Emotions.txt include stringul care desemnează emoţiile din înregistrări. Acest fişier este utilizat la verificarea corectitudinii denumirilor directoarelor destinate să conţină fişierele înregistrărilor cu emoţii, respectiv şi fişierele anexe corespunzătoare. Se observă că mai mult din fişierele de intrare folosite sunt destinate doar etapei de verificare automată a corectitudinii denumirilor şi adnotărilor. Fişierele de înregistrare au extensia .wav, fişierele de adnotare au extensia TextGrid. Erorile depistate sunt înscrise într-un fişier special numit Dc.log. Acelaşi fişier este folosit pentru înscrierea şi a altor tipuri de erori depistate în fazele următoare ale programului. Fişierul va fi util ulterior în corectarea corpusului respectiv.

Programul tratează (blocul tratarea erorilor conform figura 2) şi alte tipuri de erori existente în corpus, de exemplu incompletitudinea corpusului (lipsa unor înregistrări cu o emoţie, sau existenţa unui fişier cu valorile F0, dar nu şi cu valorile F1-F3, etc.). Erorile depistate sunt incluse în fişierul Dc.log.

Fişierul binar Dc.bin conţine toate datele de intrare (mai puţin .wav) cu structuri organizate în forma unui array cu elemente de dimensiuni egale. Principalele elemente sunt structuri - de tipuri diferite. De exemplu, structura vorbitorilor conţine informaţiile din fişa vorbitorului. Prin această structurare, informaţia este accesibilă uşor pentru prelucrări statistice. Rezultatele programului sunt fişierele text Dump.txt şi Statistics.txt (vezi figura 2).

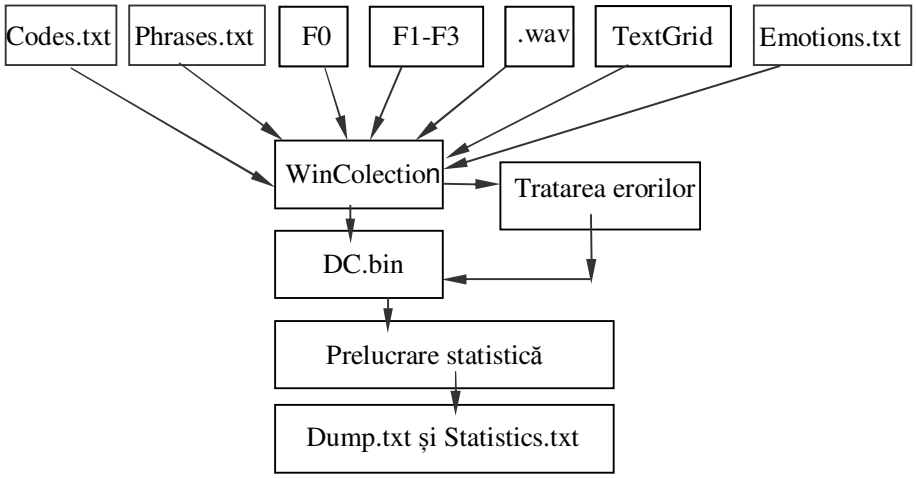


Figura 2: Organigrama programului WinCollection

Programul WinCollection creează colecţia de date, fişierul binar DC.bin, care se va obţine după parcurgerea şi procesarea tuturor fişierelor de intrare. Acest fişier este creat în directorul colecţiei de date, fiind ulterior folosit de aplicaţie în analiza statistică a vocilor emoţionale.

3.1 Etapele programului

Etapele programului sunt:

1. încărcarea informaţiilor despre vorbitori din fişierul Codes.txt;
2. parcurgerea fişierului Phrases.txt şi încărcarea informaţiilor despre propoziţiile pentru care există înregistrări;
3. încărcarea informaţiilor despre stările emoţionale pentru care există înregistrări, din fişierul Emotions.txt;
4. procesarea datelor de intrare în vederea detectării şi semnalării eventualelor erori pe baza informaţiilor din fişierele Codes.txt, Phrases.txt şi Emotions.txt;
   - 4.1. încărcarea în memorie a datelor de intrare din fişierul de adnotare şi a fişierelor cu valorile formanţilor F0 şi F1-F3;
   - 4.2. semnalarea în fişierul Dc.log a paşilor încărcării datelor şi semnalarea erorilor;
5. crearea fişierul binar Dc.bin, care va conţine întreaga colecţie de date;
6. crearea fişierelor text, Dump.txt şi Statistics.txt, descrise ulterior.

Programul parcurge toate sub-directoarele care conţin înregistrări despre o anumită stare emoţională, fiecare stare emoţională având în corespondenţă un sub-director care poartă numele respectivei stări. În fiecare sub-director, programul identifică toate seturile de câte patru fişiere (fişierul .wav, adnotarea corespunzătoare împreună cu fişierele ce conţin valorile formanţilor F0, F1-F4). Identificarea se face pe baza codului vorbitorului şi a propoziţiilor pentru care există înregistrări.

3.2 Fişiere de ieşire

Fişierul binar conţine şi următoarele informaţii suplimentare: data şi ora creării, versiunea de program utilizată, informaţii despre persoana care a realizat prelucrarea, etc.

Fişierul text Dump.txt conţine:

- informaţii generale despre colecţie (data creării, versiunea colecţiei, versiunea programului WinCollection, etc.);
- un sumar al bazei de date (număr vorbitori, fraze, stări, număr înregistrări pentru fiecare stare şi frază);
- datele de intrare;
- un set de rezultate statistice pentru fiecare vocală, inclusiv precizarea numărului de apariţii pentru fiecare stare emoţională (rezultatele se refera la minimul, maximul, valoarea medie şi dispersia, calculate pentru valorile F0 respectiv F1-F3, pentru fiecare vocală, pentru fiecare apariţie a vocalei într-o înregistrare).

Fişierul Statistics.txt generat de către programul WinCollection conţine următoarele informaţii:

- pentru fiecare vocală, numărul de apariţii pentru fiecare stare emoţională;
- pentru fiecare vocală, valoarea medie pentru populaţie, pentru formantul F0, respectiv F1-F3;

3.3 Tratarea erorilor

Programul semnalează, în urma prelucrării datelor de intrare, erorile găsite. Principalele erori semnalate de către program în urma prelucrării datelor sunt:

- lipsa adnotării pentru o înregistrare existentă în directorul analizat;
- erori în adnotare (apariţia la un anumit nivel a unui fonem care nu apare în propoziţia înregistrată);
- absenţa fişierului care conţine informaţiile despre formantul F0 pentru o anumită înregistrare;
- absenţa fişierului care conţine informaţiile despre formanţii F1-F4 pentru o anumită înregistrare.

După corecţia erorilor, este necesară re-procesarea tuturor datelor de intrare în vederea obţinerii fişierului binar. Nu se fac adăugiri la colecţie; adăugarea unei noi înregistrări la o colecţie deja existentă se face prin re-procesarea datelor de intrare şi obţinerea unui nou fişier binar.

În prezent analiza statistică este implementată pe o vocală specifică, pentru un acelaşi vorbitor, în cadrul aceleiaşi propoziţii specifice, pentru o stare specificată.

3.4 Estimarea variabilităţii

Exemplificăm utilizarea programului în vederea obţinerii unui coeficient de variabilitate inter şi intra-stare pentru valorile formanţilor precum şi o statistică generală a formanţilor pentru vocale în limba română („a", „e", „i", „u", „ă") în funcţie de starea emoţională. În cadrul analizei statistice, primul autor a introdus un coeficient de variabilitate (asimetrie) definit conform formulei:

$$\eta = \frac{1}{4}\sum_{k=0}^{3}\frac{1}{N}\sum_{\varphi=1}^{N}\frac{1}{M}\sum_{\Psi=1}^{M}\frac{|F_k[j;\varphi,\psi]-F_k[\psi]|}{F_k[\psi]}, \qquad (1)$$

unde: M- este numărul de apariţii ale fonemului ψ în fraza φ respectivă, iar k – numărul formantului curent. $F_k[\psi]$ reprezintă valoarea medie a formantului k, pentru fonemul ψ, aşa cum este determinată pentru un mare număr de vorbitori, în diverse contexte de pronunţie.

Se recunoaşte imediat că se realizează media diferenţelor absolute ale valorilor formanţilor respectivi, în pronunţiile date ale fonemului ψ, diferenţe normalizate cu valoarea medie a formantului respectiv pentru fonemul respectiv, $F_k[\psi]$: $\frac{1}{M}\sum_{\Psi=1}^{M}\frac{|F_k[j;\varphi,\psi]-F_k[\psi]|}{F_k[\psi]}$. Deoarece în sumă apar diferenţele unui singur vorbitor faţă de toţi ceilalţi, semnificaţia este aceea a diferenţei vorbitorului faţă de media vorbitorilor (specificitatea vorbitorului în pronunţia fonemului respectiv, în cadrul frazei date, pentru formantul ales). Aceste valori medii devin elemente în următoarea sumă, care, prin multiplicare cu 1/N devine media pentru toate frazele (mediile pe frază sunt mediate în sumă după φ). În fine, se face prin ultima sumă media, pentru toţi formanţii, a diferenţelor vorbitor- media vorbitorilor. De remarcat că dacă se calculează numai sumele interioare -care reprezintă rezultate parţiale în program – se determină specificitatea vorbitorului la nivel de formant al unui fonem în fraza specificată, respectiv la nivel de formant al fonemului specificat pe toate frazele. Desigur, pentru fiecare fonem (aici, vocală) şi pentru fiecare vorbitor, se obţine câte un coeficient de variabilitate.

„Coeficientul de asimetrie" în pronunţie a vorbitorului v_j faţă de media generală la aceeaşi stare şi aceeaşi vocală, pentru formantul k este dat de formula:

$$\alpha_{B,v,k} = \frac{\overline{F}_{\#k}(starea\,"B";vocala"a";vorbitor"v") - \overline{F^0}_{\#k}(B,a,toti\ vorbitorii)}{\overline{F^0}_{\#k}(B,a,toti\ vorbitorii)}, \qquad (2)$$

4. Exemple de rezultate preliminare

Exemplele prezentate nu reprezintă rezultate definitive.

Numărul de ocurenţe pentru fiecare vocală în fişierele care au fost analizate sunt date în tabelul 1. Menţionăm că vocalele analizate din cele patru propoziţii sunt opt „a", patru „e", patru „i", doi „ă" şi un „u", pentru cinci vorbitori, patru stări emoţionale cu o medie

de trei rostiri ale fiecărei propoziții. În cazul stării de tristețe am supus analizei doar două propoziții pentru trei vorbitori, iar pentru neutru trei propoziții pentru doi vorbitori.

Tabel 1: Numărul de apariții ale vocalelor analizate în funcție de starea emoțională

Vocale	Starea de bucurie	Starea de furie	Starea de tristețe	Ton neutru
A	84	103	52	102
E	40	50	21	54
I	39	48	24	60
Ă	22	27	14	24
U	11	14	0	9

Exemplificăm, în tabelele 3 și 4, utilitatea metodei expuse pentru caracterizarea unui vorbitor. Dintre cei cinci vorbitori analizați în această lucrare, unul (11861) se distinge de ceilalți prin valorile mult diferite pentru vocalele „u", „e", „i", pentru valorile formanților F0 și F1, iar alt vorbitor (05392) se distinge prin valorile mult mai mici pe vocalele „e" și „a", pentru valorile F0 și F3. Această concluzie este justificată de valorile prezentate în tabelele 3, 4 și 5. Această distanțare nu se manifestă la formanții F1 și F2 în general.

Tabel 2: Valorile coeficientului de variabilitate pentru toate fonemele („a", „e", „i", „u", „ă") analizate, pentru toți vorbitorii analizați în funcție de starea emoțională

Codul Vorbitorului	Starea de bucurie	Starea de furie	Starea de tristețe	Tonul neutru
11861	0.156	0.150	0.110	0.178
12121	0.157	0.172	0.102	0.126
83714	**0.206**	**0.166**	**0.116**	**0.219**
05392	**0.093**	**0.091**	**0.104**	**0.059**
32167	-	0.102	0.049	0.103

Tabel 3: Valorile coeficientului de asimetrie pentru valorile formanților F0-F3 în funcție de starea emoțională, pentru vorbitor 11861m, vocala „a"

11861/vocala A	Coef. de asimetrie pentru valoarea lui F0	Coef. de asimetrie pentru valoarea lui F1	Coef. de asimetrie pentru valoarea lui F2	Coef. de asimetrie pentru valoarea lui F3
Starea de bucurie	**0.33**	**0.01**	0.03	0.003
Starea de tristețe	**0.39**	0.18	0.28	**0.10**
Starea de furie	**0.27**	**0.03**	0.12	**0.01**
Ton neutru	**0.33**	**0.10**	0.13	**0.07**

Tabel 4: Valorile coeficientului de asimetrie pentru valorile formanților F0-F3 în funcție de starea emoțională, pentru vorbitor 05392m, vocala „a"

05392/vocala A	Coef. de asimetrie pentru valoarea lui F0	Coef. de asimetrie pentru valoarea lui F1	Coef. de asimetrie pentru valoarea lui F2	Coef. de asimetrie pentru valoarea lui F3
Starea de bucurie	0.11	0.11	0.18	0.07
Starea de tristețe	0.11	0.16	0.12	0.03

05392/vocala A	Coef. de asimetrie pentru valoarea lui F0	Coef. de asimetrie pentru valoarea lui F1	Coef. de asimetrie pentru valoarea lui F2	Coef. de asimetrie pentru valoarea lui F3
Starea de furie	0.09	0.05	0.12	0.04
Ton neutru	0.08	0.13	0.10	0.02

Tabel 5: Valorile coeficientului de asimetrie pentru valorile formanţilor F0-F3 în starea de bucurie, pentru patru vorbitori, vocala „u”

Vorbitori/vocala U/Starea de bucurie	Coef. de asimetrie pentru valoarea lui F0	Coef. de asimetrie pentru valoarea lui F1	Coef. de asimetrie pentru valoarea lui F2	Coef. de asimetrie pentru valoarea lui F3
11861	0.38	0.32	0.31	0.11
12121	-	0.47	0.19	0.16
83714	0.23	0.43	0.29	0.15
05392	0.13	0.24	0.12	0.07

Rezultatele din tabelul 2 arată că valorile coeficienţilor de variabilitate sunt mai mari în cazul stării de bucurie şi sunt mai scăzute în cazul stării de tristeţe. În tabelele 3, 4 şi 5 valorile coeficientului de asimetrie sunt cele mai mari în cazul frecvenţei fundamentale, iar cele mai mici pentru formantul de ordin trei.

S-a urmărit diferenţierea unui vorbitor faţă de ceilalţi în funcţie de stările emoţionale (tabel 2) precum şi analiza parametrilor F0-F3 pentru toţi vorbitorii, în funcţie de vocale (tabel 5). În tabelele 3 şi 5 am exemplificat diferenţele ce apar între doi vorbitori pentru aceeaşi vocală „a” pe baza parametrilor F0-F3 în funcţie de stările emoţionale.

5. Concluzii şi direcţii viitoare

Scopul principal al acestei lucrări a fost de a prezenta metodologia de organizare a datelor în vederea procesării lor cu o nouă aplicaţie informatică dezvoltată de colectivul nostru şi de a prezenta sumar capabilităţile aplicaţiei. Am evidenţiat principalele obiective care au stat la baza metodologiei de organizare şi avantajele aplicaţiei realizate. S-au prezentat etapele obţinerii colecţiei de date precum şi posibilitatea de manipulare pentru prelucrări statistice. Analiza sumară realizată exemplifică flexibilitatea în utilizare.

În viitor, pe baza informaţiilor privind sexul vorbitorului din fişierul Codes.txt se vor adapta limitele acceptabile pentru F0 ce se vor aplica în validarea valorilor din fişierul cu F0, notat F0 în figura 2. În etapa următoare, analiza va fi extinsă la toţi vorbitorii din cadrul corpusului SRoL.

Mulţumiri. Cercetarea a fost realizată cu sprijinul Academiei Române, în cadrul temei interne a Institutului de Informatică Teoretică din Iaşi. Autorii mulţumesc celorlalţi co-autori ai sitului Sunetele Limbii Române pentru includerea altor înregistrări noi care au fost analizate în lucrare precum şi referenţilor care au realizat observaţii pertinente.

Contribuţia autorilor: Primul autor a iniţiat tema cercetării şi structura lucrării, a elaborat metodologia generală de lucru, conceptul şi structura generală a aplicaţiei informatice, a propus formulele menţionate

în lucrare și a coordonat întreaga cercetare. Al doilea autor a implementat în C++ programul WinCollections. Al treilea autor a realizat înregistrările, a efectuat manual adnotările și a creat cu utilitarul Praat fișierele necesare utilizării programului WinCollections. Toți autorii au contribuit la redactarea lucrării.

Referințe bibliografice

Boersma, P., Weenink, D., Institute of Phonetic Science, University of Amsterdam, Praat: doing phonetics by computer, *www.praat.org*.

Feraru, M., Teodorescu, H.N. (2009). Classification of the Emotional States in Speech using the SRoL Database – Preliminary Results, *Proc. 3rd Int. Conf. Electronics, Computers and Artificial Intelligence*, Pitesti, România, ISBN 1843-2115, 27-32.

Kienast, M., Sendlmeier, W. F. (2008). Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech, An Acoustic Framework for Detecting Fatigue in Speech Based Human-Computer-Interaction, *Lecture Notes in Computer Science*, ISBN 978-3-540-70539-0, 5105/2008, 54-61.

Mcgilloway, S., Cowie, R., Douglas-Cowie, E. et al. (2000). Approaching Automatic Recognition of Emotion from Voice: A rough Benchmark. *Proc. ISCA Workshop Speech and Emotion, Newcastle*, 207-212.

Nakatsu, R., Solomides, A., Tosa, N. (1999). Emotion Recognition and its Application to Computer Agents with Spontaneous Interactive Capabilities. *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, Florence, Italy, 2, 804-808.

Scherer, K., A. (2000). Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology. *Proc. Conf. Spoken Language Processing (ICSLP)*, China, http://www.unige.ch/fapse/emotion/publications/pdf/icspl00_crosscul.pdf.

Teodorescu, H.N., Feraru, M. (2007). A study on Speech with Manifest Emotions, 10th International Conference on Text, Speech and Dialogue, *Lecture Notes in Computer Science, Springer Verlag*, ISBN 978-3-540-74627-0, 4629/2007, 254-262.

Teodorescu, H.N., Feraru, M., Trandabăț, D., Zbancioc, M., Luca, R., Verbuță, A., M. Ganea, R. Voroneanu, O. Pistol, L. (2005). Proiectul Sunetele Limbii Române, *www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm.*

Ververidis, D., Kotropoulos C. (2006). Emotional speech recognition: Resources, features, and methods, *Speech Communications*, 48: 9, 1162-1181.

PROGRAM DE EDITARE CONTURI INTONAȚIONALE ÎN LIMBA ROMÂNĂ BAZAT PE IERARHII DE FORME PROSODICE FUNCȚIONALE

VASILE APOPEI, DOINA JITCĂ, OTILIA PĂDURARU

Institutul de Informatică Teoretică

Academia Română - Filiala Iaşi

vapopei@ iit.tuiasi.ro

Rezumat

În lucrare se prezintă un program-editor de arbori intonaționali ce poate fi utilizat în analiza şi sinteza intonației în limba română. Editarea unui contur intonațional este legată de un text de intrare (frază) şi implică împărțirea acestuia în unități prosodice pe mai multe nivele. Editorul ajută la structurarea unităților în arbori de rostire, precum şi la asocierea lor cu funcții şi forme de contur F0. Formele aparțin unui set de categorii prosodice funcționale la nivelul actului comunicativ, definite în cadrul modelului de intonație pe care se bazează editorul. În secțiunea 2 se prezintă modelul de intonație, sub aspectul perspectivei ierarhice şi funcționale asupra structurii conturului intonațional. În secțiunea 3 se prezentă funcționarea editorului pentru construirea arborilor, pentru gestionarea fişierelor conținând arbori şi pentru transformarea arborilor în contururi intonaționale prin sinteză vocală.

1. Introducere

Lucrarea prezintă un program de editare grafică-interactivă a contururilor intonaționale în limba română bazat pe un set de categorii predefinite de forme elementare pentru conturul frecvenței F0. Un astfel de program ar putea fi folosit în învățarea unor elemente noi de modelare a intonației în limba română, cât şi în învățarea limbii române ca limbă străină, pentru antrenarea cursanților în formarea unor intonații adecvate diferitelor tipuri de enunțuri. Contururile intonaționale pot fi construite din forme predefinite, funcție de tipul şi structura enunțului. Acest tip de program poate fi folosit de specialiştii din domeniul analizei şi sintezei vocale dar şi în domeniul învățării limbii române, ca limbă străină, fiind un instrument de antrenare pentru producerea pattern-urilor intonaționale în mod conştient. În cadrul unor astfel de activități, studenții vor trebui să învețe semnificația diferitelor pattern-uri intonaționale în cadrul discursului şi apoi să le recunoască la nivel perceptual [Rocca, (2007)].

În proiectarea programului s-a folosit experiența acumulată în proiectarea modulului de generare a conturului F0, în sistemul text-voce pentru limba română, dezvoltat la Institutul de Informatică Teoretică [Apopei et al., (2007)], [Jitca et al., (2009a)].

Modalitatea de descriere a contururilor intonaționale pe baza unor forme elementare de contur F0, a fost posibilă după identificarea unor categorii funcționale de unități prozodice elementare şi a studierii combinațiilor acestora la nivelul unităților melodice neelementare. Definirea categoriilor funcționale s-a bazat pe stabilirea unei relații între conturul F0 al unităților elementare şi funcția acestora în realizarea actului comunicativ.

În secțiunea 2 se prezintă modelul de intonație, cu aprofundare asupra ierarhiei unităților intonaționale şi a realizării contururilor melodice ca secvențe ierarhizate ale

acestor unităţi. Cunoaşterea acestor aspecte este necesară pentru înţelegerea modului de operare cu programul de editare. În secţiunea 3 sunt prezentate funcţiile editorului pentru generarea unei intonaţii în limba română, precum şi managementul structurilor arborescente intonaţionale.

2. Modelul de intonaţie

Modelul intonaţional aflat la baza editorului face parte din categoria celor perceptuale ca şi modelul IPO [J.'t Hart et.al., (1990)], care idealizează un contur F0 natural prin aproximarea cu o secvenţă de forme elementare selectate dintr-un set predefinit. Formele de contur F0 corespund unităţilor prosodice din care se consideră a fi compus conturul F0 analizat. Modelul de intonaţie pentru limba română, prezentat în [Jitca et al.(2009b)] cuprinde două aspecte importante ce vor fi prezentate în subsecţiunile următoare. Subsecţiunea 2.1 prezintă tipurile de domenii prosodice şi realizarea lor prin unităţi în cadrul arborelui unei rostiri. Subsecţiunea 2.2 prezintă categoriile funcţionale în baza cărora se face împărţirea unităţilor prosodice, astfel încât conturile F0 neelementare să poată fi descrise prin secvenţe de unităţi funcţionale.

2. 1. Perspectiva ierarhică a structurii conturului intonaţional

Pentru a da un înţeles contururilor F0 complexe, s-a procedat la împărţirea acestora în segmente melodice elementare (SME), identificarea pattern-urilor lor melodice şi a raporturilor tonale între acestea în cadrul segmentelor melodice neelementare (nSME). SME-urile sunt segmente care includ un singur eveniment fonologic de tip accent şi corespund unităţilor de accentuare (în engl. „*accentual unit*"-AU) definite în cadrul modelelor de intonaţie autosegmenntal-metrice. La nivel textual, AU-rile corespund unui cuvânt cu accent, sau unui cuvânt însoţit de eventuale cuvinte neaccentuate.
SME-urile prin caracteristica lor tonală (ton ţintă sau nivel tonal mediu) intră în contraste tonale cu alte SME-uri din vecinătate sugerând formarea unor grupuri (grupuri de unităţi de accentuare - în engl. „*accentual uni group*" - AUG) cărora le corespund segmentele melodice neelementare (nSME). Noi considerăm că un grup este, în esenţă, un contrast între doi „poli" tonali şi că nu este obligatorie prezenţa tonurilor de graniţă pentru delimitarea lor. Caracteristica tonală a grupului este dată de nivelul ţintă dominant sau nivelul tonal mediu, în lipsa unuia ţintă dominant. Prin caracteristica tonală dominantă grupurile pot intra în relaţie de contrast tonal cu elementare vecine ce sunt observabile la un nivel de grupare superior.
Cu ale cuvinte o frază intonaţională se poate descompune intr-o ierarhie de segmente elementare şi neelementare. Spre exemplu, o rostire neutrală a textului *Lui Winston / (i se treziră / (nişte vagi /amintiri))* poate fi descrisă prin unităţile demarcate pe text prin „/" şi paranteze rotunde iar în ierarhia din figura 1 prin unităţile AU1-AU4 la nivel elementar şi AUG1-AUG2 la nivel neelementar.

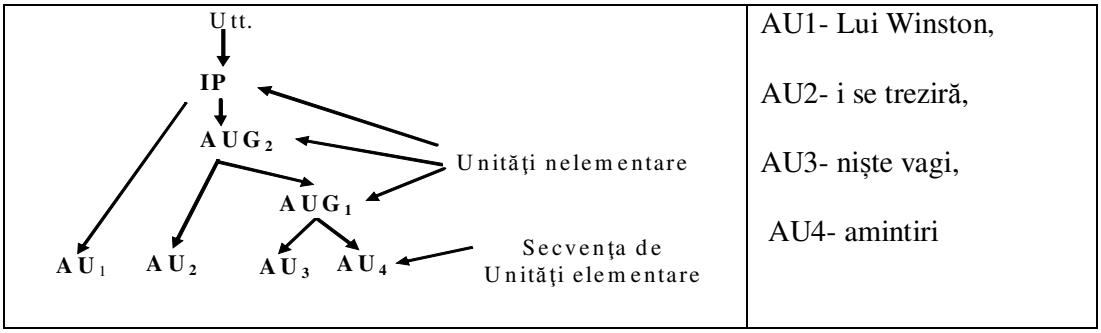


Figura 1: Arborele intonaţional al unei rostiri neutrale a textului
„Lui Winston / (i se treziră / (nişte vagi /amintiri))"

Modalitatea prin care se proiectează diferitele nivele ale unităţilor din cadrul arborelui intonaţional în cadrul rostirii, fie în varianta naturală sau sintetizată, o constituie gama de variaţie a conturului F0 în cadrul acestora (unităţilor de nivel mai înalt le corespund game mai largi ale variaţiei frecvenţei F0). Nivelele arborelui din figura 1 sunt mai bine puse în evidenţă de reprezentarea echivalentă din figura 2. Pe frunzele arborelui sunt unităţile de tip AU iar în noduri cele de tip grup de AU (AUG). Ierarhia modelului poate conţine un nivel corespunzător frazei intonaţionale (Nivel I), două nivele de AUG (Nivel II şi Nivel III) iar pe ultimul nivel sunt situate unităţile prosodice elementare (AU).

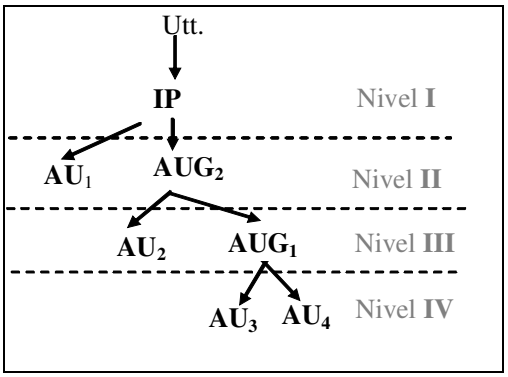


Figura 2: Arborele intonaţional al rostirii configurat ca secvenţe funcţionale pe mai multe nivele

Conturul F0 este redat în sinteză pin concatenarea formelor de contur F0 ale unităţilor elementare (AU) după poziţionarea lor în spaţiul (timp, frecvenţă F0), determinată atât de structura arborelui intonaţional cât şi de funcţiile acestora în cadrul grupurilor din care fac parte.

2.2. Perspectiva funcţională asupra conturului F0

Pattern-urile contururilor intonaţionale elementare au fost inventariate în urma analizei unui corpus de voce în limba română şi au fost împărţite pe categorii funcţionale la nivelul actului comunicativ. Categoriile au fost asociate unor etichete pentru a fi folosite în descrierea contururilor F0 [Jitca et. al.(2008)]. Perspectiva modelului nostru asupra intonaţiei echivalează din punct de vedere funcţional, unităţile elementare cu cele

neelementare, şi ca urmare, funcţiile sunt aceleaşi cu cele de la nivelul unităţilor elementare, prezentate în tabelul 1.

Tabelul 1: Semnificaţia etichetelor funcţionale aplicate unităţilor intonaţionale

Eticheta	Funcţie comunicativă		Etichetă	Funcţie comunicativă
PH	PUSH – „împinge" înainte actul de comunicare prin tonuri înalte care atrag atenţia ascultătorului		PH+f	Realizează un eveniment PUSH şi o focalizare neutrală minoră
PO	POP – relaxează actul comunicativ fie pentru închiderea unităţii (caz afirmativ) fie pentru începerea unităţilor din finalul interogaţiilor de tip Da/NU		PH+F	Realizează un eveniment PUSH şi o focalizare neutrală majoră
PD	PUSH-DOWN – efectuează un eveniment de tip PUSH şi apoi aduce tonul la nivel jos (sil. Neacc.)		PD+F	efectuează un eveniment de tip PUSH şi apoi o focalizare neneutrală (sil. acc.)
PU	POP-UP . - efectuează un eveniment de tip POP şi apoi ridică tonul sugerând continuarea		PO+f	Realizează un eveniment POP şi o focalizare neutrală minoră
f	*Focus minor*- realizează focalizare neutrală minoră (cuvintele funcţionale la nivel gramatical)		PO+F	Realizează un eveniment POP şi o focalizare neutrală majoră
F	*Focus major*-realizează focalizare neutrală majoră (cuvintele nefuncţionale la nivel gramatical)		L	LINK – realizează o funcţie de legătură între două unităţi funcţionale contrastante

Aceste categorii de funcţii sunt implementate la nivelul modulului de sinteză vocală prin seturi de prototipuri de contururi F0 elementare, care generează la nivelul percepţiei segmente melodice elementare.

3. Prezentarea programului de editare

Programul facilitează construirea structurilor intonaţionale ierarhice ale rostirilor (arborii rostirilor – în engleză, *utterance tree*) prin manevrarea unor obiecte vizuale corespunzătoare segmentelor elementare şi neelementare, până la realizarea unui contur intonaţional dorit corespunzător rostirii unui text în limba română. Arborii de rostiri constituie intrările/ieşirile unui astfel de editor asupra cărora se pot aplica următoarele operaţii:

- **Creare -** construire arbore nou
- **Editare** – modificarea arborelui curent dintr-un fişier XML
- **Salvare** - pentru memorarea arborelui editat într-un fişier text de tip XML, prin înlocuirea celui iniţial sau prin adăugare.
- **Vizualizare** – pentru afişarea grafică a arborilor intonaţionali conţinuţi într-un fişier XML.
- **Redare** – pentru a comanda sinteza vocală a rostirii pe baza a arborelui intonaţional curent şi a textului conţinut de acesta.

Pentru operaţia de **salvare** a arborilor generaţi prin comenzile **creare/editare** s-au folosit fişiere XML în care textul corespunzător arborelui intonaţional este adnotat cu un set de tag-uri şi atribute prezentate în tabelul 2.

Tabelul 2: Tag-uri şi atribute pentru adnotarea prosodică a unui text

Tag	Atribut	Valoare	Tipul unităţii
<IP>			Frază intonaţională
<AUG>	*Function*	PH, PO, PU, PD, PH+f, PO+f, PD+F, PH+F, PO+F,f , F, L	Grup de unităţi de accentuare
	TonalContrast	s(trong),w(eak)	
<AU>	*Function*	PH, PO, PU, PD, PH+f, PO+f, PD+F, PH+F, PO+F,f , F, L	Unitate de accentuare
	PichAccent	H*, L*, L+H*, H+!H*, L*+H	
	TonalContrast	s(trong),w(eak)	
<Syl>	*Length*	Small, Medium, Large	Silabă
	Energy	Small, Medium, Large	

Folosind tag-urile, cu atributele şi valorile din tabelul 2, arborii intonaţionali creaţi de editor, ca ierarhii de obiecte, corespunzătoare unităţilor prozodice, ce includ textul aferent, se transformă în urma comenzii de **Salvare** în structuri XML de text adnotat prozodic.

Operaţia de **Creare/Editare** a arborilor implică definirea în memorie a unor obiecte de tip „*segment contur intonaţional*" corespunzătoare unităţilor prozodice (unităţi de accentuare şi grupări ale acestora). Aceste obiecte au câte un echivalent vizual în fereastra principală, de editare arbore, ce face posibilă manevrarea lor cu mouse-ul (figura3).

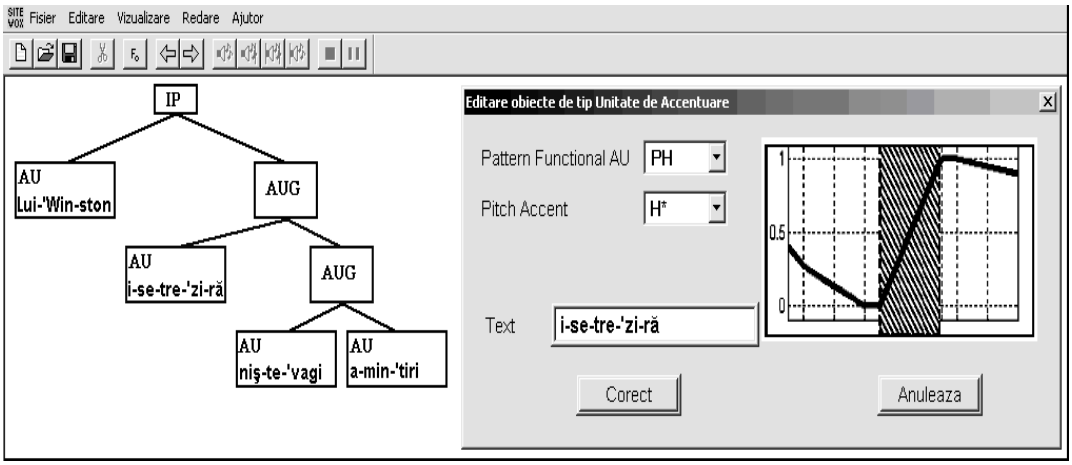


Figura 3: Fereastră de editare a unui arbore intonaţional

Comenzile de bază aplicate obiectelor vizuale pentru construirea unui arbore sunt:

- **Selecţie –** selecţie obiect/grup obiecte
- **Inserare** – inserare unitate AU
- **Ştergere** – ştergere Unitate AU/AUG
- **Grupare** – grupare unităţi AU/AUG selectate în vederea creării unui nou AUG

Comenzile de inserare, ştergere se aplică în raport cu obiectul selectat. Comanda de grupare se aplică obiectelor adiacente selectate, aparţinând aceluiaşi nivel.

Fiecare obiect are un set de atribute dintre care funcţia unităţii la nivelul actului comunicativ este cel mai important pentru că acesta este folosit în selectarea pattern-ului de contur F0 şi/sau poziţionarea acestuia în gama frecvenţei F0. Aplicarea valorilor pentru atribute şi editarea textului aferent unităţilor de accentuare se face cu comanda **Editare** în cadrul unei ferestre de dialog - figura 3.

4. Concluzii

Programul de editare de arbori intonaţionali este un instrument util pentru generarea de intonaţii corecte în limba română corespunzătoare diferitelor structuri gramaticale şi sintactice. Pentru construirea unui arbore care să genereze o intonaţie validă operatorul trebuie să cunoască modelul intonaţional bazat pe ierarhii de forme prosodice funcţionale. Editorul va putea fi folosit în proiectarea de reguli pentru modulul de predicţie prozodică. În cazul folosirii metodelor de învăţare automată se pot construi baze paralele de arbori intonaţionali şi sintactici pe baza cărora să se poată deduce reguli de punere în corespondenţă.
Odată cu implementarea unui model de intonaţie pentru o anumită limbă a apărut ca o aplicaţie imediată instruirea studenţilor străini în învăţarea intonaţiei limbii respective. Dar instrumentul poate fi folosit chiar şi pentru a instrui vorbitorii nativi ai limbii respective în a gândi şi înţelege un contur intonaţional prin prisma elementelor modelului.

Referinţe bibliografice

Apopei, V., Jitcă, D. (2007). Module for generating the F0 Contour using as input a Text structured by prosodic information, *Advances in Spoken Language Technology, Romanian Academy, 119-126*.

Rocca, A. P. D. (2007). New Trends on the Teaching of Intonation of Foreign Languages, *Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech,* 420-428.

J.'t Hart, Collier R., Cohen A. (2006). A perceptual study of intonation, Cambridge University Press.

Jitcă, D., Apopei V. (2009a). A Prosodic Control module for a Romanian TtS System, based on melodic contour dictionaries, *Advances in Spoken Language Technology, Romanian Academy,* 77-86.

Jitcă, D., Apopei, V. and Jitcă, M., (2009b). The F0 Contour Modelling as Functional Accentual Unit Sequences*, International Journal of Speech Technology, DOI 10.1007/s10772-009-9055-3, Nov, 2009*.

Jitcă, D., Apopei V. (2007). Corpus de voce pentru limba română adnotat cu etichete funcţionale la nivelul unităţilor de accentuare, *Lucrările atelierului „Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, p.31-39, Iaşi, 15.11.2007.

CORPUS PENTRU GNATOFONIE: PROTOCOL, METODOLOGIE, ADNOTARE

HORIA-NICOLAI TEODORESCU[1,2], ALINA UNTU[1]

[1]*Universitatea Tehnică „Gheorghe Asachi", Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei, Iaşi - România*
[2] *Institutul de Informatică Teoretică, Academia Română, Iaşi - România;*

{hteodor, auntu}@etti.tuiasi.ro

Rezumat

Prezentăm elemente privind extinderea unui micro-corpus de sunete gnatofonice, exemple de analiză formantică şi temporală a consoanelor fricative în cazuri de protezare şi edentaţie, precum şi rezultatele preliminare.

1. Introducere

Gnatofonia a fost introdusă de către primul autor, parţial la sugestia Prof. Leonid Teodorescu şi reprezintă o metodă de analiză a deficienţelor de pronunţie a sunetelor datorate diverselor patologii ale aparatului stomatognat (edentaţie, malocluzie, disfuncţii ale articulaţiei temporo-mandibulare, etc.). Primul micro-corpus de sunete gnatofonice (înregistrate pe casetă) a fost realizat de către primul autor în colaborare cu C. Morărașu, V. Burlui şi D. Leca şi a stat la baza primelor validări semiempirice în gnatofonie, raportate intern în teza de absolvire a C. Morărașu, iar apoi, cu unele dezvoltări datorate în principal lui C. Morărașu, într-un set de lucrări comunicate sau publicate în anii 1990 (Burlui, Teodorescu, Morărașu, 1993 a, b, 1994, 1996). Un alt corpus care a fost introdus pe sit-ul „Sunetele limbii române" la secţiunea „Arhiva pentru aplicaţii de gnatosonie şi gnatofonie", descris în (Teodorescu et al., 2005-2007), (Teodorescu et al., 2007), (Teodorescu, Feraru, 2007 a) a fost realizat de primul autor în colaborare cu S.M. Feraru şi a fost inclus în (Teodorescu, Feraru, 2007 b). Dacă termenul de gnatofonie şi conceptul de disciplină au fost introduse în anii 1980, observarea empirică a modificărilor fonatorii induse de edentaţie sunt cunoscute încă din anul 1959 când s-au realizat primele studii sistematice de către Rathbone şi Snidecor (Rathbone, Snidecor, 1959). Autorii au studiat instrumental diferenţele ce apar în pronunţia de propoziţii test ce conţin consoane post-dentale (*n*, *t*, *d*, *r*, *l*, *s*, *sh*, *z*, *zh*, *y*), linguo-dentale (*th* sonor şi nesonor), labio-dentale (*f*, *v*) şi combinaţii post-dentale (*ch*, *j*) înainte şi după tratamentul ortodontic a 10 pacienţi cu diferite stadii de malocluzie. Au constatat că înainte de tratament din 16 sunete dentale testate un procent de 6.4 % sunt defectuoase, cele mai mari erori concentrându-se asupra consoanelor dentale de tip fricativ (*s*, *z*, *sh*, *zh* şi *th*). După patru ani, opt pacienţi din 10 au fost reexaminaţi după aceeaşi procedură, rezultând o scădere a procentului de sunete cu pronunţie defectuoasă de până la 1.5 %. Erorile de pronunţie s-au păstrat la consoanele fricative având însă un grad mai scăzut, iar celelalte sunete dentale au fost corectate astfel încât erorile au fost indetectabile. Acest studiu a dovedit că defectele de vorbire se corectează după tratamentul ortodontic fără a fi nevoie de aplicarea unei terapii de educare a vorbirii. Analiza de semnal vocal în cazul unor patologii ale aparatului stomatognat este importantă în diagnosticarea şi evaluarea eficienţei tratamentului ortodontic.

Spre sfârşitul anilor 1990, bazându-se pe un număr de 42 de articole consultate, Johnson şi Sandy (Johnson, Sandy, 1999) au prezentat stadiul cercetărilor realizate până în anul 1999 de către specialişti în domeniu, vizavi de relaţia dintre poziţia dinţilor şi vorbire. S-a constatat că dinţii au un rol important în vorbire, însă relaţia dintre poziţia dinţilor şi vorbire a rămas încă controversată. Se cunoaşte capacitatea pacienţilor de a-şi compensa defectele de vorbire ce apar în cazul malocluziilor, însă mecanismul ce stă la bază rămâne necunoscut. Deşi s-a demonstrat că există o relaţie între defectele de dentiţie şi defectele de pronunţie, nu s-a putut concluziona că ar exista o corelaţie şi cu severitatea malocluziilor.

2. Metodologia de culegere a datelor

Metodologia, protocolul de înregistrare şi protocolul de documentare au fost preluate de pe sit-ul „Sunetele limbii române". Subiecţii au fost informaţi anterior înregistrărilor de obiectivele proiectului, fiind asiguraţi de confidenţialitatea datelor personale. Subiecţii au semnat un consimţământ informat în conformitate cu protocolul de protecţie a subiecţilor umani şi cu principiile etice ale cercetărilor care implică fiinţa umană existente la nivel naţional şi internaţional. Cercetarea pe subiecţi umani privind analiza sunetelor gnatofonice şi gnatosonice a fost aprobată de către consiliul Facultăţii de Electronică, Telecomunicaţii şi Tehnologia Informaţiei din cadrul Universităţii Tehnice „Gheorghe Asachi" din Iaşi (Teodorescu et al., 2005-2007)[1].

2.1 Metodologia de înregistrare

Sunetele gnatofonice au fost înregistrate cu ajutorul unui microfon prevăzut cu căşti, A4 Tech Stereo HS-60, având caracteristicile: frecvenţă de răspuns 20 Hz-20 kHz, impedanţă: U=3V, R=1,5kΩ, sensibilitate: -58dB±2. Placa de bază a calculatorului pe care au fost efectuate înregistrările este Sony Vaio MBX-189 având încorporată o placă de sunet Intel® High Definition Audio compatible 3D audio (Direct Sound 3D support) cu următoarele caracteristici: procesor de semnal 44-kHz / 16-bit stereo CD quality, mod de ieşire a sunetului 8 canale, 192 kHz / 32 bit, standard Intel HD audio.

Pentru înregistrări s-a utilizat programul GoldWave^{TM}, versiunea 5.54, la o frecvenţă de eşantionare de 22050 Hz cu atributele PCM signed (16-24 bits mono). Culegerea de semnal vocal se realizează în condiţii de zgomot redus (amplitudinea zgomotului trebuie să fie mai mică cu cel puţin 20 dB decât amplitudinea frecvenţei fundamentale). Conform metodologiei de înregistrare de pe situl SRoL, se recomandă ca poziţia microfonului să fie mai jos de gură, aproximativ în dreptul bărbiei (la câţiva centimetri de aceasta), iar distanţa de la bărbie să fie aproximativ egală cu distanţa până la buze (Teodorescu et al., 2005-2007).

2.2 Metodologia de documentare

Fişele subiecţilor „Profil vorbitor", „Chestionar Patologie Vocală şi Factori Obiectivi" şi „Fişă dinţi subiect" au fost completate de către al doilea autor. Pentru „Fişă dinţi subiect" s-au luat în considerare informaţii ce pot fi cunoscute de către subiectul

[1] Avizul este accesibil la:
www.etc.tuiasi.ro/sibm/romanian_spoken_language/images/jpg/doc_subiecti_umani_analiza_gnatofonie.jpg

înregistrat sau vizualizate de către persoana care completează fișa, de tip prezență / absență dinte, plombă, implant, coroană, punte. Am considerat elemente de interes și plombele întrucât acestea dacă nu sunt realizate corect pot modifica structura dintelui având repercusiuni asupra spațiului interdentar, în consecință și asupra vorbirii (spațierile dintre incisivi pot produce în vorbire efectul de „fonfăit"). Pentru a valida aceste idei este necesară o analiză statistică pertinentă (cel puțin 20 de subiecți). Un exemplu de fișă ce conține informații referitoare la patologiile arcadelor superioară și inferioară este dat în figura 1. Subiectul prezintă opt plombe simple la nivelul maxilarului și opt plombe simple la nivelul mandibulei, care afectează primii molari de pe fiecare hemiarcadă.

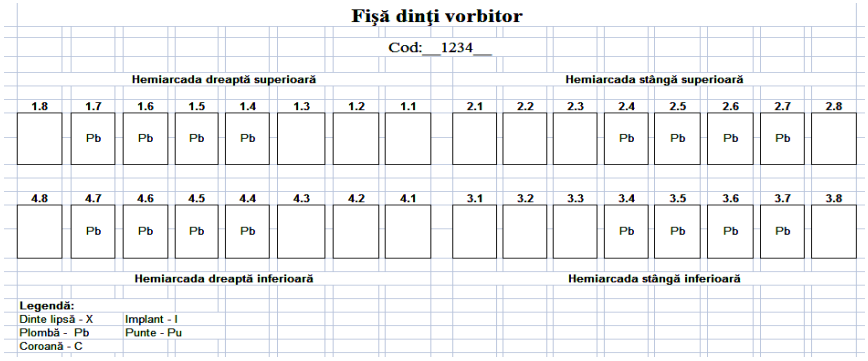


Figura 1: Exemplu de fișă dinți subiect.

2.3 Listă cuvinte

Pentru realizarea corpusului de sunete gnatofonice s-a utilizat setul de cuvinte stabilit anterior de primul autor (©), cuvinte ce conțin consoanele *s*, *f*, *ș*, *v*, *z*, *j*. S-au ales aceste cuvinte, întrucât studiile realizate până în prezent de către cercetători în domeniu au arătat că fricativele prezintă mai frecvent modificări de pronunție în cazul existenței unor patologii ale aparatului stomatognat (Rathbone, Snidecor, 1959). Analiza comparativă a pronunției consoanelor *s*, *f*, *ș*, și a celor semi-vocalice *v*, *z*, *j*, furnizează informații despre afectarea dentiției.

Cuvintele utilizate pentru înregistrările gnatofonice sunt: vată / fată, var / far, vuiet (pronunțat vvvvvuiet) / vuiet (pronunțat normal, scurt, vuiet) / fffffui / fui, vvvvvaiet / vaiet / fffffaieton / faieton, vecin / fecior, vvvvvânt / vânt / fffffân / fân, vvvvvine / vvvvvine / vvvine / fffffine, vine / fine, vehement / ferment, vierme / fierbe, vâjât / gâjât / sâsât / fâsât, bâzâie / zâzâie, bâzzzzzâie / zâzzzzzâie, fâșât / fâlfât, vâjâie / fâlfâie / sâsâie / fâșâie, vâjjjjjâie / fâlfffffâie / sâssssâie / fâșșșșșâie, vâjjjjjât / sâssssâit / fâssssâit / fâșșșșșâit / fâlfffffât, vorbit / fortuit / sortit, suit / vuit.

Scopul cercetării noastre este de a identifica și analiza modificările ce apar în dinamica formanților / pseudo-formanților consoanelor fricative pronunțate de către două clase de vorbitori: o clasă martor formată din vorbitori cu dentiție normală (neafectată de patologii ale aparatului stomatognat) și o clasă de vorbitori cu defecte de dentiție. Setul de cuvinte întocmit în acest scop include cuvinte ce conțin consoanele *v*, *f*, *s*, *ș*, *j*, *z* în context consoană-vocală (CV) sau vocală-consoană-vocală (VCV). Termenul de „pseudo-formanți" a fost folosit de primul autor în (Teodorescu, Feraru, 2008), pentru a desemna traseele asemănătoare cu cele ale formanților pe care le detectează utilitarul

Praat™ în cazul consoanelor fricative. Sunetele neperiodice cum ar fi siflantele *s*, *f*, *ş*, nu prezintă frecvenţă fundamentală, drept urmare nu au nici formanţi propriu-zişi (care au, prin definiţie, frecvenţe egale cu multipli ai frecvenţei fundamentale).

Pronunţiile celor 56 de cuvinte pentru fiecare din cei 10 vorbitori sunt înregistrate într-un singur fişier .wav care este salvat cu un nume mnemonic, în formatul „cod_subiect_sex" (ex. 1234_m). Fişierele au fost filtrate şi re-salvate în format .wav şi în format .ogg (16 şi 24 biţi). Înregistrările au fost ascultate pentru o analiză perceptuală a zgomotului, iar pronunţiile cu trunchiere (saturaţie) în amplitudine au fost eliminate.

3. Discuţia corpusului

Corpusul de sunete gnatofonice este disponibil pe sit-ul „Sunetele limbii române", în cadrul arhivei pentru aplicaţii de gnatofonie şi gnatosonie (Figura 2). În paralel am realizat şi un corpus de sunete gnatosonice pe care îl vom prezenta într-o lucrare ulterioară.

Arhiva pentru aplicaţii de gnatofonie şi gnatosonie

Metodologia de culegere a semnalelor gnatofonice este identică cu cea de culegere de semnal vocal.

Cuvintele utilizate pentru înregistrările gnatofonice sunt alese astfel încât să se poată analiza comparativ modificările de siflante, fricative şi de consoane semi-vocalice.

Cuvintele utilizate pentru înregistrările gnatofonice sunt: vată / fată; var / far; vuiet (pronunţat vvvuiet) / vuiet (pronunţat normal, scurt, vuiet) / fui / vaiet (pronunţat vvvaiet) / vaiet (pronunţat vaiet) / faieton / vecin / fecior / vânt (pronunţat vvvânt) / vânt (pronunţat vânt) / fân / vvvvine, vvine, vvine / vine / fine / vehement / ferment / vierme / fierbe / vâjât / vvvvâjjjjât / vvvâjjjjâie / ffffâşşşşşşâie / ffffâşşşşşât / fâşât / sâsât / sssssâsssssâie / gâjât / zâzâie / bâăzzzzâăâie / bâz&226;ie.

Un numar de 10 înregistrări gnatofonice provenite de la 5 subiecţi de gen feminin şi 5 subiecţi de gen masculin sunt adnotate, iar fişierele TextGrid şi wav corespunzătoare se pot vizualiza (accesa) la adresa (http://www.etc.tuiasi.ro /sibm/romanian_spoken_language/ro/sunete_gnatofonice.htm) sau aici.

Fişă vorbitor	Gnatofonie		Gnatofonie - filtrate		Gnatosonie		Gnatosonie - filtrate	
	wav	ogg	wav	ogg	wav	ogg	wav	ogg
Fişă vorbitor #394715	394715_f 394715_f_v1 394715_f_v2	394715_f 394715_f_v1 394715_f_v2	394715_f 394715_f_v1 394715_f_v2	394715_f 394715_f_v1 394715_f_v2				
Fişă vorbitor #231515	231515_m	231515_m	231515_m	231515_m				
Fişă vorbitor #280269	280269_m	280269_m	280269_m	280269_m				
Fişă vorbitor #20048					20048_f_v0	20048_f_v0	20048_f_v0	20048_f_v01
Fişă vorbitor #55555	55555_f 55555_f_v1	55555_f 55555_f_v1	55555_f 55555_f_v1	55555_f 55555_f_v1	55555_f	55555_f	55555_f	55555_f
Fişă vorbitor #1234	1234_m	1234_m	1234_m	1234_m	1234_m	1234_m	1234_m	1234_m
Fişă vorbitor #1357	1357_m	1357_m	1357_m	1357_m	1357_m	1357_m	1357_m	1357_m
Fişă vorbitor #2001	2001_f	2001_f	2001_f	2001_f	2001_f	2001_f	2001_f	2001_f

Figura 2: Arhiva pentru aplicaţii de gantofonie şi gnatosonie.

3.1 Statistica înregistrărilor şi patologiilor

Am realizat înregistrări de sunete gnatofonice provenind de la cinci subiecţi de gen masculin şi cinci subiecţi de gen feminin, cu vârste cuprinse între 21 şi 46 ani,

majoritatea cu studii superioare, fără afecțiuni respiratorii, laringeale, neurologice, sau psihologice.

Corpusul realizat are la bază înregistrări provenind de la nouă subiecți fără defecte majore de dentiție și fără deficiențe de vorbire detectabile prin percepție auditivă și un subiect cu edentație majoră (13 molari lipsă). Starea dentiției subiecților înregistrați este ilustrată în Tabelul 1.

Tabel 1: Starea dentiției subiecților înregistrați

Cod subiect	**Sex**	**Vârstă (ani)**	**Nr. Pb (plombă) / localizare**		**Nr. X (lipsă dinte) / localizare**		**Nr. C (coroană) / localizare**		**Nr. Pu (punte) / localizare**		**Tip voce / patologii**
2404	M	30	1	primul incisiv HSS	-	-	-	-	-	-	Profesională, fără patologii
2001	F	26	7	primul molar HDI, al 2-lea și al 3-lea molar HDS și HSS, al 2-lea molar HSI, al 3-lea molar HDI	-	-	1	al 2-lea molar HDI	-	-	Neprofesională, fără patologii
4312	F	29	3	primii incisivi HSS și HDS, al 3-lea molar HDS	1	al 3-lea molar HDI	2	al 2-lea molar, HSS, al 3-lea molar HSI	-	-	Profesională, fără patologii
1357	M	30	2	canin HSS, al 2-lea molar HSS	2	ultimul molar HDI și HSI	2	al doilea molar HDI și HSI	1	molarii 1, 3, HDS	Neprofesională, fără patologii
2202	F	26	3	al 2-lea molar HSS, al 3-lea molar HDS, al 4-lea molar HSI	1	al 3-lea molar HSI	-	-	-	-	Neprofesională, fără patologii
1234	M	30	16	primii 4 molari de pe fiecare hemiarcadă	-	-	-	-	-	-	Neprofesională, fără patologii
3371	F	21	4	primul incisiv HSS, al 2-lea molar HSS, al 4-lea molar HDI, al 3-lea molar HSI	2	al 3-lea molar HDI și HSI	-	-	-	-	Neprofesională, fără patologii
3298	M	30	3	2 pe primii incisivi de pe HSS și HDS, al 3-lea molar HDS	-	-	2	al 2-lea molar HDI și HSI	-	-	Profesională, fără patologii
01321	F	30	2	canin și primul molar HDI	-	-	1	primul molar HSS	-	-	Neprofesională, fără patologii
5343	M	46	1	ultimul molar HSI	13	incisivii 1, 2 HSS, primii 3 molari HDS, ultimii 3 molari HSS, molarii 2,3,4 HSI, molarii 3, 4 HDI	-	-	-	-	Neprofesională, fără patologii

În studiul nostru ne interesează în special afectarea incisivilor, caninilor și primilor molari deoarece aceștia intervin în articularea sunetelor. Cei patru subiecți la care am efectuat analiza consoanelor fricative (1234m, 5343m, 01321f, 3371f) prezintă

următoarele patologii relevante: primul subiect prezintă câte o plombă la fiecare din cei patru molari primari; subiectul al doilea prezintă edentaţie ce interesează primul şi al doilea incisiv de pe hemiarcada stângă superioară (HSS) şi primul molar de pe hemiarcada dreaptă superioară (HDS); al treilea subiect prezintă câte o plombă pe caninul şi primul molar de pe hemiarcada dreaptă inferioară (HDI) şi o coroană pe primul molar de pe HSS; cel de-al patrulea subiect prezintă o plombă pe primul incisiv de pe HSS.

4. Metodologia de analiză

4.1 Metoda de prefiltrare

Toate fişierele au fost prefiltrate cu un filtru trece bandă cu frecvenţele de tăiere de cca. 70 Hz şi 7 kHz şi cu atenuare de 100 dB la frecvenţele de 50 Hz şi 12 kHz (în afara benzii de trecere), setat în utilitarul GoldWave™ (Figura 3). Filtrul se poate aplica o dată sau, repetat, de două ori (echivalent, două filtre înseriate) astfel ca nivelul final de zgomot la 50 Hz să fie cu cel puţin 20 dB mai mic decât nivelul la frecvenţa fundamentală (F0). În cazul înregistrărilor curente, filtrul s-a aplicat o singură dată.

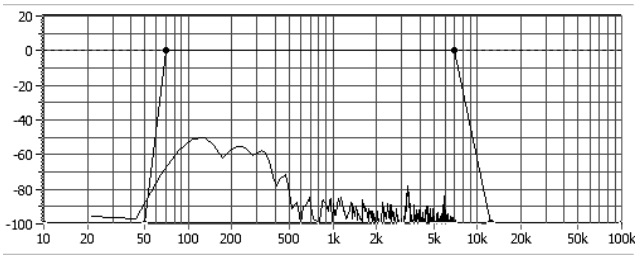


Figura 3: Filtru trece bandă de la 70 Hz la 7 kHz setat în utilitarul GoldWave™ .

4.2 Metodologia de adnotare

Pentru o analiză ulterioară a diferenţelor / similitudinilor ce apar la nivelul consoanelor fricative, am segmentat şi adnotat cele 10 fişiere la nivel de fonem / silabă / cuvânt cu ajutorul utilitarului Praat™ (Boersma, 2002). Am realizat împărţirea fişierelor în subfişiere ce conţin consoanele *f*, *s*, *ş*, *v*, *z*, *j*, în context CV şi VCV. Au rezultat câte nouă sau 11 subfişiere la fiecare subiect, numite astfel: vf_CV(a)_cod_subiect_sex – consoanele *v* şi *f* în context CV, V=vocala *a*, jsfshz_VCV(a-)_cod_subiect_sex – consoanele *j*, *s*, *f*, *ş*, *z* în context VCV, V=vocala *â*, vfs_CV(o)_cod_subiect_sex – consoanele *v*, *f*, *s* în context CV, V=vocala *o*, etc. Pentru *ş*, *â* şi *ă* am utilizat notaţiile *sh*, *a-* şi *a+*.

În cadrul procesului de adnotare am luat în considerare şi pauzele intravorbire (ce apar în rostirea de silabă, cuvânt) notate cu simbolul $, pauzele intervorbire (dintre cuvinte) fără notaţie (blanc) şi pauzele care nu se aud, notate cu simbolul %. Adnotarea s-a realizat ţinând cont de percepţia auditivă, prezenţa / absenţa frecvenţei fundamentale (ce evidenţiază caracterul vocalic, sonor, cu activarea corzilor vocale), forma de undă, energia şi modificările ce apar pe spectrogramă de la un fonem la altul.

4.3 Metodologia de analiză formantică şi temporală

Cu ajutorul utilitarului Praat™ am extras F0, formanţii / pseudo-formanţii F1, F2, F3, F4 şi timpii corespunzători fiecărui fonem şi fiecărui cuvânt în parte. Pentru extragerea formanţilor / pseudo-formanţilor F1, F2, F3, F4, Praat-ul utilizează o fereastră glisantă de analiză de lungime 0,025 s, cu un pas de deplasare de 0,00625 s, preaccentuare de 50 Hz şi frecvenţe maxime pentru formanţi de 5500 Hz pentru voci feminine şi 5000 Hz pentru voci masculine. Conform manualului de utilizare al utilitarului Praat™, preaccentuarea de 50 Hz determină nemodificarea spectrului la frecvenţe sub 50 Hz şi amplificarea cu 6 dB la fiecare octavă: 6 dB la frecvenţa de 100 Hz, 12 dB la frecvenţa de 200 Hz , etc. (Boersma, 2002).

Pentru extragerea frecvenţei fundamentale (F0), manualul Praat™ recomandă utilizarea benzilor de 75 Hz-300 Hz pentru voci masculine, respectiv 100 Hz-500 Hz pentru voci feminine, precum şi a unui pas de deplasare al ferestrei de analiză de 0,01 s. „Se folosesc intervale standard pentru F0 la voci masculine şi feminine întrucât alegerea unei valori prea mici determină pierderea modificărilor rapide ale lui F0, iar alegerea unei valori prea mari duce la pierderea valorilor foarte mici ale lui F0” (Boersma, 2002).

Analiza a constat din următoarele etape: i) Segmentarea, în vederea analizei modificărilor duratelor fonemelor. Segmentarea este efectuată la nivelurile fonem, silabă şi cuvânt, pentru a se determina momentele de timp corespunzătoare graniţelor dintre foneme şi cuvinte. Într-o fază ulterioară, vom realiza o segmentare mai fină, care să permită determinarea duratelor *glissando*-urilor fonemelor (de ex., *va*, *fa*, *rfff* – de ex. în „*până-n vârfffful muntelui”* etc.); ii) Generarea fişierelor .txt cu instrumentul Praat, cu valorile formanţilor în evoluţia lor; iii) Importarea fişierelor într-un utilitar (de ex., Excel™) şi separarea manuală a fonemelor în funcţie de limitele temporale corespunzătoare fiecărui fonem; iv) Analiza dinamicii şi analiza statistică a formanţilor pe segmentele de interes, comparativ între subiecţii cu afecţiuni şi tratamente ale aparatului stomatognat şi subiecţii sănătoşi. Am analizat dinamica formanţilor / pseudo-formanţilor consoanelor *v*, *f*, *s* din cuvintele *vată*, *fată*, *sâsâit*, *fâsâit*, *sâsssâit*, *fâsssâit*. Acestea s-au aplicat la patru vorbitori (doi de gen feminin şi doi de gen masculin).

5. Exemple de rezultate preliminare

Studiul realizat pe baza corpusului de sunete gnatofonice s-a axat pe analiza defectelor de pronunţie cauzate de patologii ale aparatului stomatognat. S-a urmărit efectuarea unei comparaţii între dinamica formanţilor consoanelor fricative la subiecţi cu dentiţie normală şi subiecţi cu defecte de dentiţie.

În continuare prezentăm rezultate preliminare obţinute în urma procesului de adnotare şi a analizei formanţilor / pseudo-formanţilor consoanelor *v* şi *f* din cuvintele *vată*, *fată*. Graficele cu traseele F1, F2, F3, pentru toate cuvintele extrase spre analiză au fost efectuate pentru toţi cei patru subiecţi amintiţi anterior. Un exemplu de adnotare a unui fişier ce conţine consoanele siflantă *f* şi semi-vocală *v* în context CV este ilustrat în figura 4. Partea superioară reprezintă semnalul în domeniul amplitudine-timp („forma de undă”), partea mediană reprezintă traseele formanţilor (pseudo-formanţilor) F1, F2, F3, F4, iar partea inferioară adnotarea la nivel de fonem (1), silabă (2) şi cuvânt (3).

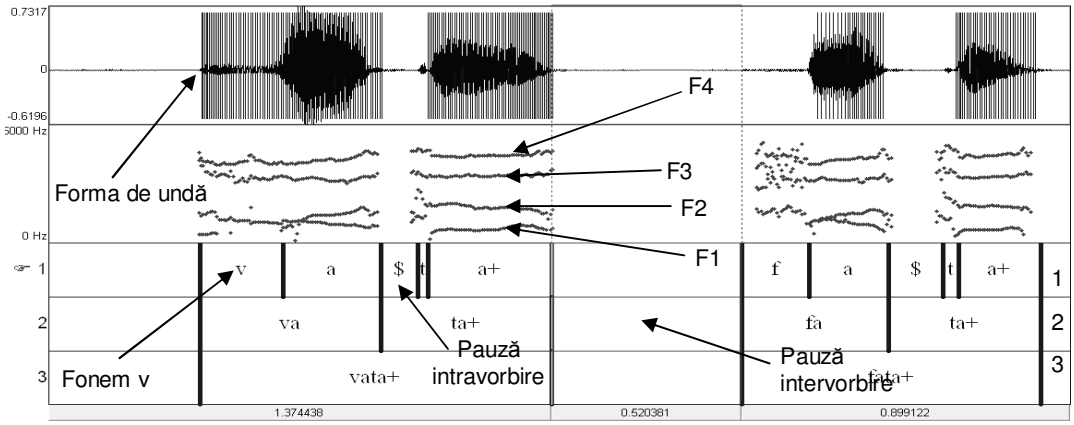


Figura 4: Exemplu de adnotare în utilitarul Praat™ (Vorbitor 1234m).

În graficul a) din figura 5 sunt reprezentate traseele formanților (pseudo-formanților) F1 și F2 ai consoanelor *v* și *f* în context CV (cuvintele *vată*, *fată*), pentru subiectul cu edentație majoră 5343m. La consoana *v*, se observă că formantul F1 are valori de cca. 250 Hz și formantul F2 de cca. 1000 Hz, la aproximativ prima jumătate a traseului, ulterior valorile acestora modificându-se brusc la 1000 Hz respectiv 2000 Hz.

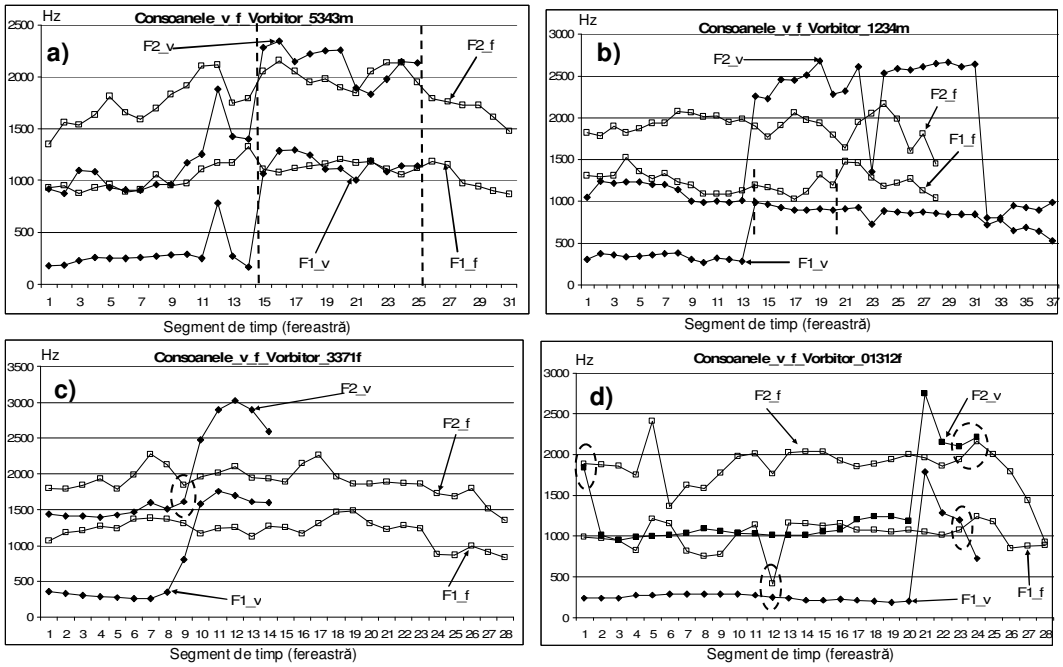


Figura 5: Traseele formanților / pseudo-formanților F1, F2, pentru consoanele *v* și *f* din cuvintele *vată*, *fată*, pentru patru vorbitori -5343m, 1234m, 3371f, 01312f ce prezintă patologiile amintite anterior (vezi cap. „Statistica patologiilor").

Se constată că valorile pseudo-formanților consoanei *f* au valori foarte apropiate cu cele ale formanților din a doua jumătate a traseelor consoanei *v*. Prin urmare consoana *v* alunecă prin valorile formanților către un *f*. Această alunecare poate fi interpretată ca o deficiență în pronunția consoanei *v*, dar o analiză de detaliu este necesară pentru a confirma această concluzie preliminară. În graficele b), c) și d) sunt reprezentate

traseele formanţilor (pseudo-formanţilor) consoanelor *v* şi *f* pentru subiecţii 1234m, 3371f şi 01312f, care nu au defecte majore de dentiţie. Comparativ cu subiectul ce prezintă edentaţie, se observă că la ceilalţi trei subiecţi traseele formanţilor F1 şi F2 pentru consoanele *v* şi *f* diferă între ele cu excepţia zonelor marcate cu linii punctate, care cuprind: un eşantion pentru formantul F2 la subiectul 3371f, două eşantioane pentru F1 respectiv trei eşantioane pentru F2 la subiectul 01312f şi şapte eşantioane pentru formantul F1 la subiectul 1234m.

6. Concluzii şi direcţii viitoare

Rezultatele obţinute indică o posibilă corelare între modificările de dentiţie şi caracteristicile formantice la pronunţia consoanelor fricative. Având în vedere că sunt foarte preliminare este necesară o bază statistică mai mare pentru a deriva concluzii solide. Analiza gnatofonică se află încă în stadiul de cercetare, dar în viitor poate constitui un instrument util în stomatologie. Efectuarea de înregistrări gnatofonice pe pacienţi cu defecte de dentiţie de tip edentaţie, malocluzie, disfuncţii ale articulaţiei temporo-mandibulare, înainte de aplicarea tratamentului, poate avea relevanţă în stabilirea diagnosticului şi evaluarea deficienţei de pronunţie indusă de patologiile aparatului stomatognat. Corpusul de sunete gnatofonice poate fi utilizat şi în alte aplicaţii medicale cum ar fi detectarea defectelor de pronunţie care apar în cazul unor patologii respiratorii, laringeale, neurologice, psihologice, în măsura în care va include şi astfel de patologii. O altă aplicaţie constă în stabilirea eficienţei tratamentului logopedic, prin analiza articulării cuvintelor înainte şi după aplicarea acestuia. Ca direcţie viitoare ne propunem să dezvoltăm corpusul de sunete gnatofonice adnotate, prin realizarea de înregistrări atât pe subiecţi cu patologii ale aparatului stomatognat, cât şi pe subiecţi cu dentiţie normală.

Contribuţia autorilor. Primul autor a conceput structura şi conţinutul lucrării, metoda de înregistrare şi preprocesare a înregistrărilor, metodologia de analiză formantică şi temporală, a dictat parte din informatiile prezentate etc. Al doilea autor a realizat înregistrările folosite aici, pe care le-a filtrat, împărţit în subfişiere şi adnotat, a completat fişele subiecţilor, a extras valorile temporale şi ale formanţilor, a realizat graficele cu traseele formanţilor pentru consoanele v, f din cuvintele analizate pentru patru vorbitori. Ambii autori au contribuit la redactarea lucrării.

Mulţumiri. Autorii sunt recunoscători persoanelor care au acceptat să fie înregistrate şi au fost de acord cu utilizarea înregistrărilor în scop de cercetare şi includerea acestora în baza de date existentă pe sit-ul „Sunetele limbii române", în condiţiile protecţiei şi confidenţialităţii datelor personale. Mulţumim întregului colectiv SRoL şi în mod special colegelor M. Feraru şi L. Pistol pentru contribuţiile aduse la „Arhiva de sunete gnatofonice şi gnatosonice" de pe sit, şi colegului M. Zbancioc pentru observaţiile pertinente.

Referinţe bibliografice

Boersma, P.P.G. (2002). Praat, A System for Doing Phonetics by Computer. Glot International, Vol. 5 No. 9/10, 341-345., http://www.fon.hum.uva.nl/praat/ , data accesării: 1.04.2010.

Burlui, V., Teodorescu, H.N., Morăraşu C.S. (1993 a). L'analyse en fréquence de la fonction phonétique chez l'édenté total, *7ème Symposium Européen Sur Le Traitement de l'Edéntation Totale*, Lyon, France, 1993 (vol. de rezumate al Simpozionului).

Burlui, V., Teodorescu, H.N, Morărașu, C.S. (1993 b). L'analyse en fréquence de la restauration de la fonction phonétique chez l'édenté total, *Symposium Européen Sur Le Traitement de l'Edéntation Totale* , Lyon, Franța.

Burlui, V., Teodorescu, H.N., Morărașu, C.S. (1994). La fonction phonatoire chez l'édenté total. Analyse en fréquence. Les Cahiers de Prothèse (France), no. 88, Decembre 1994, pp. 63-68.

Burlui, V., Teodorescu, H.N., Morărașu, C. (1996). Analiza în frecvența gnatoprotetică, Sesiunea jubiliară „30 de ani de învățământ stomatologic ieșean", l martie 1996.

Johnson, N.C.L., Sandy, J.R. (1999). Tooth position and speech—is there a relationship?, *The Angle Orthodontist*, vol. 69, Issue 4, 306-310, August 1999.

Rathbone, J.S., Snidecor, J.C. (1959). Appraisal of Speech Defects In Dental Anomalies With Reference To Speech Improvement, *The Angle Orthodontist*, vol. 29, no. 1, 54-59, 1959.

Teodorescu, H.N., Burlui, V., Leca, P.D. (1986). Gnathosonic analyzer, *Med Biol Eng Comput*. July, 1988, 26(4):428-31.

Teodorescu, H.N., Feraru, S.M., Trandabăț, D., Zbancioc, M., Luca, R., Verbuță, A., Hnatiuc, M., Ganea, R., Voroneanu, O., Pistol, L., Șcheianu, D. (2005-2007), situl Web Sunetele Limbii Române. http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.html

Teodorescu, H.N., Trandabăț, D., Feraru, S.M, Zbancioc, M., Luca, R. (2007). A corpus of the sounds in the Romanian spoken language for language-related education, Chapter Six, pp. 73-90. În volumul Carlos Periñán Pascual (Editor), „Revisiting Language Learning Resources", Cambridge Scholars Publishing (CSP), UK, 2007.

Teodorescu, H.N., Feraru, S.M. (2007 a). A study on Speech with Manifest Emotions, *10th International Conference on Text, Speech and Dialogue*, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, *Lecture Notes in Computer Science*, Springer Verlag, vol. 4629/2007, 254-262.

Teodorescu, H.N, Feraru, S.M. (2007 b). Micro-corpus de sunete gnatosonice și gnatofonice, Pistol, Cristea, Tufiș (Eds.), *Resurse lingvistice și instrumente pentru prelucrarea limbii române*, Editura Universității „Al. I. Cuza" Iași, 21-30, 2007.

Teodorescu, H.N., Feraru, S.M. (2008). Classification in Gnathophonics - Preliminary Results, The Second International Symposium on Electrical and Electronics Engineering, ISEEE, 12-13 septembrie, Galați, România, Galați University Press, 525-530, 2008.

DISPOZITIV ELECTRONIC DE PREPROCESARE ÎN REGIM PARALEL A SPECTRULUI VOCAL

HULEA MIRCEA[1], UNTU ALINA[2]

[1]*Universitatea Tehnică „Gheorghe Asachi”, Facultatea de Automatică şi Calculatoare Iaşi – România;*
[2] *Universitatea Tehnică „Gheorghe Asachi”, Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei, Iaşi – România;*

mhulea@tuiasi.ro, auntu@etti.tuiasi.ro

Rezumat

Lucrarea de faţă prezintă un dispozitiv hardware bazat pe amplificatoare operaţionale a cărui funcţie principală reprezintă discriminarea frecvenţelor spectrului vocal în funcţie de formanţii specifici vocalelor limbii române. Proiectarea acestui sistem s-a realizat în scopul generării simultane în regim paralel a trenurilor de impulsuri ce stimulează o reţea neuronală analogică utilizată în domeniul recunoaşterii vocale independente de vorbitor. Astfel, prototipul prezentat în această lucrare poate fi utilizat până în acest moment pentru evidenţierea canalelor de frecvenţă ce semnalizează recepţia vocalelor. Având în vedere rezultatele obţinute în acest sens, cercetările ulterioare vor avea ca scop testarea principiilor de funcţionare ale acestui sistem şi pentru evidenţierea canalelor de frecvenţă activate la recepţia consoanelor.

1. Introducere

Un domeniu de o importanţă majoră pe plan ştiinţific reprezintă procesarea limbajului natural de către structurile de calcul artificiale. Punctul de plecare al activităţilor desfăşurate în această direcţie de cercetare constituie recunoaşterea vocală a cuvintelor care se poate realiza în mod dependent sau în mod independent de vorbitor. Cea de-a doua modalitate de recunoaştere vocală constă în extragerea şi apoi recunoaşterea unui set de caracteristici spectrale specifice informaţiei transmise prin vorbire. Extragerea proprietăţilor ce conduc la recunoaşterea vocală independentă de vorbitor a cuvintelor poate fi realizată la nivel software prin elaborarea de algoritmi sau la nivel hardware prin proiectarea de circuite electronice. Scopul acestei lucrări constă în prezentarea principiilor structurale şi funcţionale ale unui dispozitiv hardware ce realizează preprocesarea semnalului audio în vederea recunoaşterii vocalelor limbii române. Structura prototipului a fost elaborată în vederea utilizării acestuia ca element de intrare pentru reţelele neuronale analogice ce au ca unitate funcţională neuronul electronic de inspiraţie biologică. Funcţionarea acestui sistem are la bază principiile de percepţie a sunetelor evidenţiate la nivelul urechii umane care realizează divizarea spectrului audio pe canale de frecvenţă (Stevens, 1999). În această direcţie, literatura de specialitate prezintă aspecte privitoare la existenţa unor elemente de preprocesare care au fost proiectate ţinând cont de caracteristicile limbii engleze (Hopfield, Brody, 2001), (Wills, 2004), (Loizou, Dorman, 2006). Sistemul prezentat în cadrul acestei lucrări păstrează principiul de divizare pe canale de frecvenţă a spectrului vocal cu deosebirea că, filtrarea semnalului audio se realizează la nivel hardware folosind un set de parametri ce

permite captarea caracteristicilor spectrale ale vocalelor limbii române. În faza inițială, cercetările proprii efectuate în acest domeniu au folosit un element de procesare a semnalului ce avea la bază filtre rezonante trece-bandă. Performanțele acestora în delimitarea canalelor de frecvență au fost reduse deoarece prezentau o atenuare scăzută a frecvențelor din afara benzii de trecere. Efectul acestui dezavantaj crește invers proporțional cu frecvența de rezonanță a filtrelor care la valori mai mici de 1 kHz necesitau creșterea substanțială a capacităților folosite. Mai mult decât atât, pentru creșterea factorului de calitate a etajelor de filtrare și păstrarea valorilor condensatoarelor în limite acceptabile, a fost necesară utilizarea de bobine speciale cu raportul L/R ridicat. De asemenea folosirea filtrelor rezonante pentru procesarea în regim paralel a semnalului audio are ca efect negativ inductanța mutuală dintre bobine.

Prin urmare, în vederea eliminării dezavantajelor induse de filtrele rezonante, pentru obținerea canalelor de frecvență sistemul actual folosește filtre active de tip Chebyshev a căror ieșire este adaptată pentru stimularea rețelelor neuronale analogice. Frecvențele de tăiere ale acestor filtre trece-bandă s-au calculat prin utilizarea rezultatelor obținute pentru formanții F1 și F2 ai vocalelor limbii române, publicate de către H.N. Teodorescu și M. Feraru pe sit-ul „Sunetele limbii române" (SRoL) prezentat în (Teodorescu et al., 2005-2007), (Teodorescu et al., 2007), (Teodorescu, Feraru, 2007). Dintre aceste date statistice s-au folosit pentru această lucrare valorile medii ale formanților pentru vocale izolate sau în context, extrase de către autori cu utilitarul Praat™.

În continuare, secțiunea a doua a lucrării va oferi detalii privitoare la caracteristicile funcționale ale modulelor sistemului, precum și la modul de efectuare a statisticii ce a avut ca rezultat obținerea parametrilor elementelor de filtrare a semnalului vocal. Rezultatele obținute în urma testării comportamentului sistemului în cazul rostirii susținute a vocalelor sunt evidențiate în secțiunea a treia, iar concluziile referitoare la performanțele sistemului sunt prezentate în secțiunea a patra.

2. Structura sistemului

Rețelele neuronale analogice au la bază neuroni electronici ce realizează procesarea informațiilor în regim paralel folosind trenuri de impulsuri. Atât frecvența acestor impulsuri cât și energia generată de fiecare impuls pot varia în funcție de gradul de antrenare a rețelei. Pentru ca datele primite din exterior să poată fi procesate cu ajutorul structurilor neuronale de acest tip este necesară conversia în trenuri de impulsuri a semnalelor cu variație continuă. De asemenea, datorită faptului că neuronul artificial de inspirație biologică este prin esență un detector de coincidențe temporale, un interes deosebit trebuie acordat și momentelor apariției stimulilor. Astfel, rețeaua neuronală analogică va fi sensibilă atât la frecvența stimulilor de intrare cât și la gradul de concurență a acestora. Pentru stimularea unei astfel de rețele neuronale ce poate fi antrenată pentru recunoașterea vocală independentă de vorbitor este necesară o procesare prealabilă a semnalului audio. Figura 1 prezintă elementele funcționale ale sistemului de preprocesare a semnalului vocal recepționat de la un microfon. Elementul de preamplificare *PreAmp* realizează amplificarea și filtrarea semnalului audio cu scopul obținerii benzii de frecvențe corespunzătoare spectrului vocal. Semnalul obținut este normalizat de către modulul *DynAmp* prin adaptarea dinamică a amplitudinii

semnalului cu ajutorul circuitului MAX9756. Etapa următoare a procesării semnalului este îndeplinită de modulul *FTB* ce realizează divizarea pe canale de frecvență a spectrului vocal în funcție de formanții specifici vocalelor limbii române.

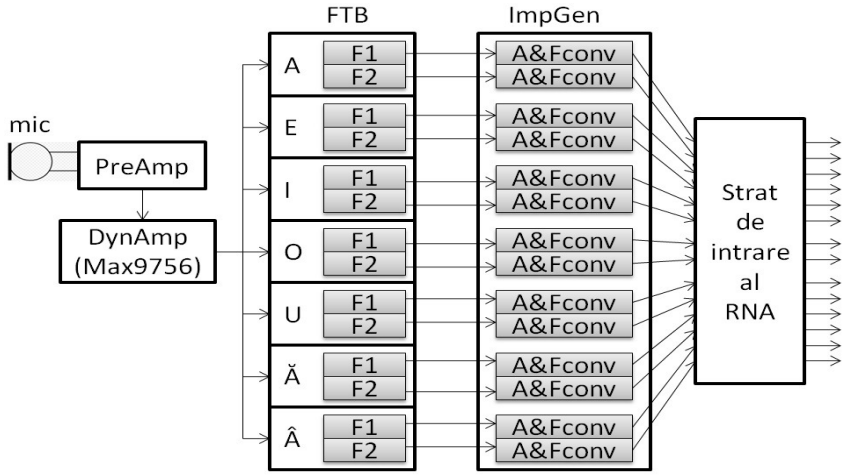


Figura 1: Structura dispozitivului de preprocesare a semnalului audio alcătuită din modulele *PreAmp* (preamplificarea semnalului), *DynAmp* (reglarea amplitudinii semnalului), *FTB* (divizarea pe canale de frecvență a spectrului vocal), *ImpGen* (conversia amplitudinii şi frecvenței semnalului); Dispozitivul stimulează stratul de intrare al rețelei neuronale analogice (RNA).

Acest modul constă într-o serie de filtre active de ordin patru ce au fost proiectate folosind amplificatoare operaționale βA741. Astfel, *F1[v]* şi *F2[v]* unde $v \in \{a, e, i, o, u, ă, â\}$ reprezintă filtre de tip Chebyshev acordate pe frecvențele specifice formanților *F1* şi respectiv *F2* ai fiecărei vocale. Pentru a fi posibilă funcționarea normală a rețelei neuronale a fost necesară proiectarea unui modul auxiliar *ImpGen* care, cu ajutorul stratului de intrare a rețelei neuronale, realizează adaptarea amplitudinii şi frecvenței impulsurilor generate de fiecare canal de frecvență al *FTB*.

2.1 Modulul DynAmp

Experimentele anterioare au evidențiat faptul că pentru obținerea unei acuități ridicate în recunoaşterea vocală a fonemelor limbii române este necesară rostirea acestora cu o intensitate ridicată şi aproximativ constantă. În vederea soluționării acestei probleme sistemul îşi propune normalizarea la nivel hardware a semnalului recepționat de la microfon după preamplificarea prealabilă a acestuia. Această adaptare de semnal se realizează cu ajutorul circuitului MAX9756 produs de firma MAXIM, ce îndeplinește funcția de reglare automată a amplitudinii semnalului. Prin utilizarea acestui circuit se reduce considerabil sensibilitatea răspunsului sistemului la variația intensității semnalului vocal de intrare. În continuare se vor prezenta parametrii utilizați în proiectarea schemei electronice de control a circuitului pentru inițializarea funcției de control a amplitudinii semnalului ALC (*Automatic Level Control).* Prin această funcție se amplifică componentele de intensitate scăzută ale semnalului fără a distorsiona componentele de amplitudine ridicată (MAXIM Innovation Delivered™, 2006). Sesizarea depăşirii unui anumit prag de putere pe ieşirea circuitului determină activarea ALC ce reduce amplificarea semnalului cu până la 6 dB. Figura 2 (a) ilustrează etapele

acțiunii ALC care sunt *creşterea* în timpul t_a a atenuării puterii semnalului, *blocarea* atenuării la valoarea curentă pe parcursul intervalului de timp t_h şi *revenirea* treptată a amplificării la valoarea normală în timpul t_r.

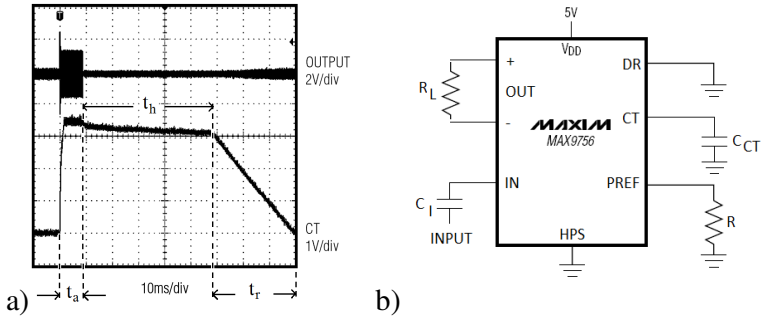


Figura 2. a) digrama de semnal obținută în urma testelor efectuate de producător asupra circuitului MAX9756 ce permite evidențierea etapelor reglării automate a amplitudinii semnalului; b) schema de conectare a pinilor prin care se definesc parametrii funcției ALC (DR, CT, PREF), precum şi a pinilor ce definesc intrarea, ieşirea şi selecția ieşirii (IN, OUT şi HPS).

Circuitul integrat oferă posibilitatea ajustării timpului de atac t_a şi a timpului de revenire t_r prin modificarea parametrilor C_{CT} şi DR. Având în vedere scopul pentru care este proiectat dispozitivul, s-a ales valoarea minimă pentru $C_{CT} = 10nF$ deoarece dezavantajele vitezei ridicate de răspuns a ALC (MAXIM Innovation Delivered™, 2006) nu sunt sesizabile în cazul recunoaşterii vocale. De asemenea, s-a ales $t_r = 300ms$ (maxim) prin conectarea *DR* la *GND* pentru a preveni fluctuațiile rapide de intensitate a semnalului vocal pe parcursul cuvintelor.

2.2 Modulul FTB

Pentru a fi posibilă utilizarea rețelelor neuronale analogice în domeniul procesării artificiale a semnalului vocal este necesar ca receptorii specializați în detecția caracteristicilor spectrale ale fonemelor să aibă capacitatea de a fi activați simultan. Acest fapt prezintă o importanță majoră datorită faptului că în mod normal un fonem este caracterizat de un număr de frecvențe a căror prezență în spectrul vocal trebuie semnalizată rețelei neuronale în timp real. Recunoaşterea fonemului respectiv se produce cu condiția ca apariția acestor frecvențe la intrarea rețelei neuronale să se producă în mod simultan. Plecând de la această idee cât şi de la modelul urechii umane, modulul *FTB* realizează divizarea pe canale de frecvență a spectrului audio. Scopul acestei abordări reprezintă extragerea în timp real a frecvențelor ce conduc la recunoaşterea informației transmise prin vorbire în mod independent de caracteristicile vocii.

2.2.1 Simularea soft a filtrelor Chebyshev

În vederea implementării etajelor de filtrare utilizate în recunoaşterea vocală independentă de vorbitor, s-au proiectat cu ajutorul utilitarului *FilterLab* oferit gratuit de *Microchip,* câte două filtre trece-bandă de tip Chebyshev pentru fiecare vocală a limbii române. Având în vedere faptul că formanții F1 şi F2 (multipli ai frecvenței

fundamentale) sunt implicaţi în discriminarea vocalelor, pentru recunoaşterea acestora s-a implementat câte un filtru ce evidenţiază recepţia fiecăruia dintre cei doi formanţi. Parametrii de proiectare utilizaţi sunt: tip filtru *Chebyshev*; riplu permis în banda de trecere -3 dB; numărul de poli 4; câştig 1. De asemenea frecvenţele de tăiere pentru cele şapte vocale (*a*, *e*, *i*, *o*, *u*, *ă*, *â*) sunt prezentate în tabelul 1.

2.2.2 Obţinerea frecvenţelor de tăiere

Pe baza valorilor medii ale formanţilor preluate de pe sit-ul SRoL, s-a calculat media şi deviaţia standard pentru formanţii F1 şi F2 specifici celor şapte vocale. Pentru stabilirea lăţimii benzii de frecvenţă a celor două filtre corespunzătoare fiecărei vocale, am ales drept frecvenţă de tăiere minimă $f_{\min}$ şi maximă $f_{\max}$ după cum urmează:

$$f_{\min} = f_{central} - st\_dev \quad (1)$$

$$f_{\max} = f_{central} + st\_dev \quad (2)$$

unde $f_{central}$ este frecvenţa centrală a benzii de trecere, iar st_dev este deviaţia standard. Atât frecvenţa centrală cât şi deviaţia standard s-au obţinut în urma statisticii efectuate pe un număr de persoane aflate în diverse stări emoţionale.
Analiza statistică s-a efectuat în mod diferit pentru grupurile de vocale *a, e, i* şi respectiv, *o, u, ă, â* deoarece am avut disponibile valori ale formanţilor specifice vocalelor rostite în contexte diferite. Pentru *a, e, i* au existat valori pentru vocale izolate, iar pentru *o, u, ă, â* au existat valori pentru vocale în context. Cele din urmă au fost extrase din propoziţiile „*Aseară*" şi „*Vine mama*" pentru patru stări emoţionale: neutru, fericire, furie, tristeţe. Frecvenţa centrală a filtrelor s-a obţinut prin calculul mediilor a 30 de valori pentru vocalele izolate *a, e, i,* iar pentru vocalele în context *o, u, ă, â,* media a fost calculată folosind 40 de valori. La ultimul grup de vocale s-au considerat 10 vorbitori (cinci de gen feminin şi cinci de gen masculin) pe patru stări emoţionale. Rezultatele obţinute sunt ilustrate în tabelul 1.

Tabel 1: Valorile frecvenţelor medii, minime şi maxime corespunzătoare formanţilor vocalelor

Vocală	**a**		**e**		**i**		**o**		**u**		**ă**		**â**	
Formant	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
$f_{central}$	751	1274	583	1889	399	2328	466	915	403	788	542	1433	389	1648
$f_{\min}$	618	1093	486	1671	338	2091	361	794	333	708	432	1192	325	1428
$f_{\max}$	883	1456	680	2107	460	2565	570	1036	473	868	651	1674	453	1867

Schema generală a unui filtru trece bandă de tip Chebyshev alcătuit din două amplificatoare operaţionale este dată în figura 3.

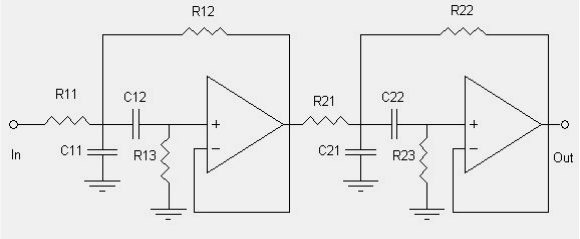


Figura 3: Schema generală a unui filtru trece bandă Chebyshev.

Pentru evidenţierea comportamentului canalelor de frecvenţă implementate, în figura 4 este prezentată simularea în *FilterLab* a răspunsului în frecvenţă a filtrului trece-bandă pentru formantul F2 al vocalei *e*. Frecvenţele de tăiere corespunzătoare la -3dB sunt $f_{\min} = 1671Hz$ şi $f_{\max} = 2107Hz$, iar lăţimea benzii de trecere este egală cu dublul deviaţiei standard

$$\Delta f_b = 2 \cdot st\_dev = 436Hz \tag{3}$$

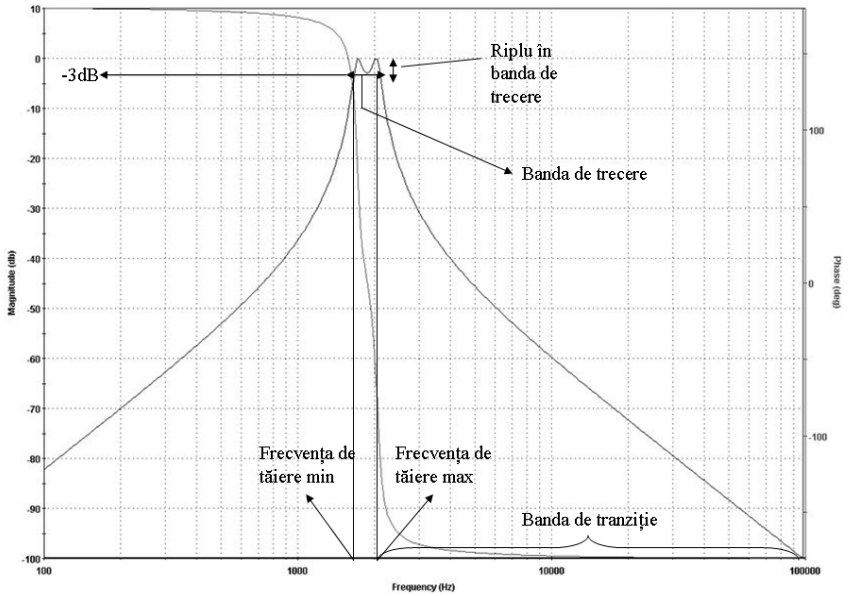


Figura 4: Răspunsul în frecvenţă a filtrului trece-bandă de tip Chebyshev.

Filtrele Chebyshev de tipul I sunt filtre polinomiale (numai cu poli, fără zerouri), având o caracteristică de modul cu ripluri egale în banda de trecere şi monoton descrescătoare în banda de oprire. Dintre toate filtrele polinomiale de ordinul *N*, filtrele Chebyshev de tipul I au zona de tranziţie cea mai îngustă. Performanţele filtrului sunt complet precizate de mărimea riplului în banda de trecere şi de ordinul filtrului ce determină panta caracteristicii de modul. Utilizând programul *FilterLab* se pot proiecta trei tipuri de filtre: Bessel, Butterworth şi Chebyshev. Filtrul Bessel se poate realiza doar pentru tipul de răspuns trece-jos, iar celelalte două pentru răspunsuri de tip trece-jos, trece-sus, şi trece-bandă. În comparaţie cu filtrul Butterworth, filtrul Chebyshev realizează o mai bună atenuare a frecvenţelor din afara benzii de trecere, acest fapt justificând alegerea făcută în procesul de proiectare a filtrelor.

2.2.3 Delimitarea canalelor de frecvență

Semnalul sinusoidal generat de filtrele trece-bandă prezentate anterior este convertit în impulsuri prin utilizarea unui comparator LM324 pentru fiecare formant. Acest etaj permite și reglarea amplitudinii de la care activitatea canalului de frecvență va produce stimularea rețelei neuronale. Prin urmare, ieșirea generată de modulul *FTB* se materializează într-un tren de impulsuri de amplitudine, durată și frecvență variabile ca în figura 5(a). Energia acestor impulsuri este convertită în trenuri de impulsuri de amplitudine și frecvență constantă de către modulul *ImpGen* ce constituie o interfață între modulul de filtrare și stratul de intrare al rețelei neuronale.

2.3 Modulul ImpGen

Experimentele realizate pe parcursul cercetărilor efectuate în domeniul recunoașterii vocale au arătat că pentru creșterea stabilității în funcționare a rețelei neuronale analogice este necesară o adaptare a amplitudinii și frecvenței impulsurilor ce stimulează intrările acesteia (Hulea, 2009). Schema electrică a acestui modul conectată cu un neuron electronic specific stratului de intrare al rețelei este prezentată în figura 5 (b).

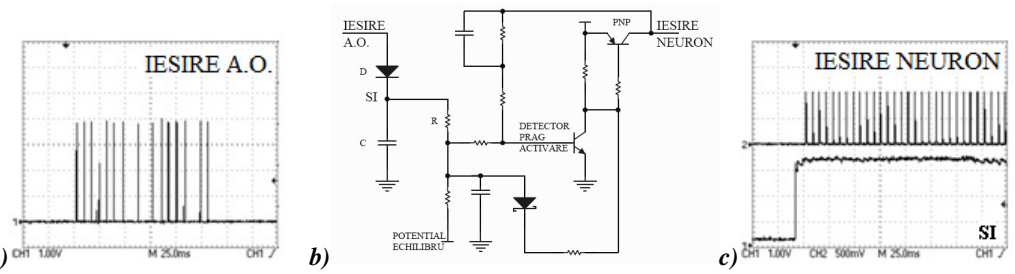


Figura 5. a) semnalul de ieșire a unui canal de frecvență al modulului *FTB*; b) schema electrică a ansamblului [modul *ImpGen* - neuron de intrare] pentru un canal de frecvență oarecare; c) ieșirea neuronului de intrare (semnalul IEȘIRE NEURON) și integrarea ieșirilor modulului *FTB* de către capacitatea *C* (Semnalul SI).

Conversia ieșirii modulului *FTB* constă în integrarea stimulilor generați și transformarea energiei obținute în impulsuri de amplitudine și frecvență constante. Prima etapă se realizează folosind capacitatea *C* ce se încarcă prin dioda *D,* iar cea de-a doua este asigurată de componentele electronice ce compun neuronul de intrare (Hulea, 2008). Frecvența impulsurilor de ieșire este invers proporțională cu valoarea rezistenței *R* prin care are loc descărcarea condensatorului *C*. În cazul sistemului descris în această lucrare, valoarea de 470 kΩ a acestei rezistențe a fost aleasă pentru a obține la ieșire o frecvență de aproximativ 160 Hz.

3. Rezultate experimentale

Răspunsurile în frecvență ale filtrelor de tip Chebyshev obținute în urma simulărilor efectuate în programul *FilterLab* au arătat că performanța acestui tip de filtru se

îmbunătățește proporțional cu ordinul acestuia. Prin urmare, odată cu creșterea numărului de etaje de filtrare utilizate, crește și panta caracteristicii de modul a filtrului. Pe de altă parte, panta maximă de tranziție se obține prin delimitarea strictă a spectrului util cu ajutorul comparatoarelor LM324 implementate pe ieșirea fiecărui canal de frecvență. Stabilitatea limitelor benzii de trecere obținute în acest mod este asigurată în cazul în care variația amplitudinii semnalului este redusă la minim. Această condiție este îndeplinită prin reglarea dinamică a amplitudinii semnalului de către modulul *DynAmp* conectat pe intrarea modulului *FTB*. Pentru ilustrarea comportamentului descris anterior, figura 6 prezintă câteva aspecte legate de funcționarea canalului de frecvență centrală 542 Hz corespunzător formantului F1 al vocalei /*ă*/. Formele de undă prezentate au fost obținute în urma evidențierii cu ajutorul osciloscopului a semnalelor generate prin nodurile de conectare a celor patru module ale sistemului. Vocalele au fost rostite independent de context la o distanță de aproximativ $3cm$ de microfon.

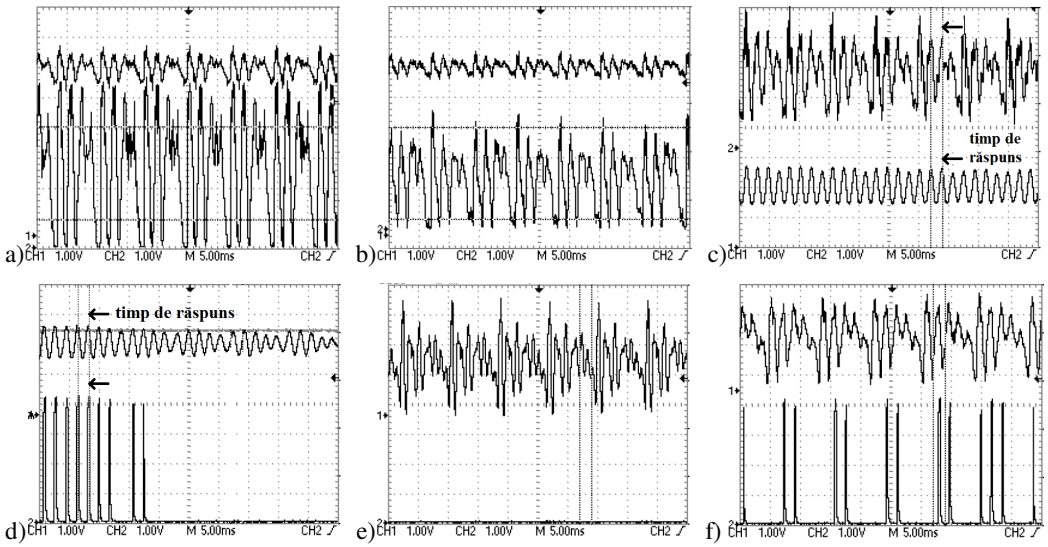


Figura 6. a) Intrarea (sus) și ieșirea (jos) a modulului *DynAmp* când funcția de reglare automată a amplitudinii semnalului (ALC) este dezactivată; b) intrarea (sus) și ieșirea (jos) a modulului *DynAmp* după activarea ALC; c) ieșirea *DynAmp* (sus) și ieșirea filtrului Chebyshev corespunzător formantului F1 (jos) în timpul rostirii vocalei susținute a vocalei /*ă*/ (cursorii verticali evidențiază timpul de răspuns a *FTB*); d) Semnalul generat de filtrul Chebyshev (sus) și ieșirea modulului *FTB* corespunzătoare formantului F1 (jos) a vocalei /*ă*/ la rostirea vocalei /*ă*/; semnalul continuu reprezintă pragul de la care semnalul generat de canalul de frecvență este considerat util (cursorii verticali evidențiază timpul de răspuns a *ImpGen*); e) ieșirea circuitului MAX9756 la rostirea vocalei /*a*/ și ieșirea FTB corespunzătoare formantului F1 al vocalei /*ă*/ f) semnalul generat de *DynAmp* la rostirea vocalei /*ă*/ și ieșirea FTB corespunzătoare formantului F1 al vocalei /*ă*/.

Așadar primele două diagrame de semnal (a) și (b) din figura 6, evidențiază importanța funcției ALC a circuitului MAX9756 în adaptarea amplitudinii semnalului generat de modulul *PreAmp*. Semnalele prezentate în figura 6 (c) prezintă efectul produs de filtrul Chebyshev de frecvență centrală 542 Hz corespunzător formantului F1 al vocalei */ă/* în momentul rostirii susținute a acesteia. Diagrama de semnal (d) ale aceleiași figuri prezintă pragul de la care activitatea canalului de frecvență va fi considerată utilă. În

continuare sunt prezentate rezultatele unui test efectuat asupra elementelor de filtrare corespunzătoare canalului de frecvenţă menţionat. Pentru efectuarea acestui test s-a rostit vocala */a/* şi s-a verificat activitatea ieşirii *FTB* corespunzătoare formantului F1 al vocalei */ă/*. După cum se poate observa, rezultatele testului au fost cele aşteptate în sensul că vocala */a/* nu a produs activarea canalului de 542 Hz (diagrama de semnal (e)), acesta fiind activat la recepţia vocalei */ă/* (diagrama de semnal (f)).

Valorile obţinute pentru frecvenţelor de tăiere ale filtrelor (tabelul 1) au impus efectuarea unor ajustări în implementarea dispozitivului pentru eliminarea elementelor de filtrare redundante. În acest sens, pentru formantul F1 al vocalelor *i,u,* şi *â* s-a utilizat acelaşi filtru de frecvenţă centrală 400 Hz, iar pentru formantul F1 al vocalelor *e* şi *ă* s-a utilizat un filtru de frecvenţă centrală 560 Hz. Diferenţa de aproximativ 200 Hz dintre valorile medii obţinute pentru formantul F2 fac posibilă discriminarea vocalelor limbii române folosind doar filtrele corespunzătoare acestui formant.

Un alt aspect important ce a fost luat în considerare în procesul de testare a sistemului reprezintă evaluarea timpului de răspuns a acestuia. Cursorii verticali de pe diagramele de semnal (c) şi (d) ale figurii 6 evidenţiază timpii de răspuns a submodulelor FTB şi respectiv *ImpGen* care sunt neglijabili în comparaţie cu perioada semnalelor specifice spectrului vocal. De asemenea diagramele de semnal (a) şi (b) ale aceleiaşi figuri evidenţiază cu uşurinţă timpul de răspuns a modulului *DynAmp* care de asemenea este neglijabil. Prin urmare se poate deduce faptul că timpul total de răspuns a întregului sistem de preprocesare este neglijabil în comparaţie cu variaţia semnalelor specifice spectrului vocal.

4. Concluzii

Pentru a beneficia de avantajele utilizării reţelelor neuronale cu răspuns în timp real în domeniul recunoaşterii vocale, s-a proiectat un sistem de preprocesare în regim paralel a semnalului audio. Plecând de la modelul aparatului auditiv uman ce realizează divizarea spectrului vocal pe canale de frecvenţă s-a implementat un set de filtre trece-bandă acordate pe frecvenţele caracteristice formanţilor vocalelor limbii române. În vederea obţinerii unei acuităţi ridicate în recunoaşterea vocală independentă de vorbitor s-a urmărit creşterea performanţelor sistemului de filtrare. Aceasta s-a realizat atât prin utilizarea filtrelor de tip Chebyshev care prezintă cea mai bună atenuare a frecvenţelor din afara benzii de trecere, cât şi prin folosirea circuitului MAX9756 care realizează adaptarea dinamică a amplitudinii semnalului. De asemenea, procesul de selecţie a frecvenţelor de tăiere pentru aceste filtre s-a bazat pe date statistice ce cuprind o serie de voci de gen masculin şi feminin precum şi diferite stări emoţionale.

Experimentele efectuate pentru studiul funcţionării sistemului prezentat în această lucrare au arătat o superioritate evidentă a performanţelor filtrelor Chebyshev în comparaţie cu filtrele rezonante utilizate în cadrul cercetărilor anterioare descrise în (Hulea, 2008). Performanţele sistemului s-au îmbunătăţit substanţial mai ales în cazul canalelor de frecvenţe centrale mai mici de 1000 Hz. Un alt avantaj al filtrelor Chebyshev reprezintă creşterea eficienţei filtrelor prin introducerea elementelor active (A.O.) şi eliminarea componentelor inductive (L). De asemenea capacităţile folosite s-au redus cu trei ordine de mărime ceea ce reduce costurile implementării. De asemenea, îmbunătăţirea performanţelor în filtrare a fost obţinută fără a afecta răspunsul în timp

real a sistemului. Totuşi o consecinţă a utilizării circuitului MAX9756 pentru adaptarea amplitudinii semnalului audio reprezintă creşterea consumului de curent al dispozitivului cu aproximativ 48 mA. Cu toate acestea, acest aspect trebuie luat în considerare doar în cazul utilizării unei surse portabile de energie pentru alimentarea sistemului. Având în vedere creşterea acuităţii în delimitarea benzilor de frecvenţă a sistemului de filtrare se preconizează pe de o parte îmbunătăţirea performanţelor în domeniul recunoaşterii vocale independente de vorbitor, cât şi simplificarea prin reducerea numărului de neuroni a arhitecturii reţelei neuronale utilizate în detecţia cuvintelor.

Referinţe bibliografice

Hopfield, J. J., Brody C.D. (2001). What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration, *Proceedings of the National Academy of Sciences of the USA*, 1282 – 1287.

Hulea, M. (2008). A Model of Silicon Neurons Suitable for Speech Recognition, Computer Engineering and Applied Informatics, Vol.10, Nr. 4, 32-41, ISSN 1454-8658.

Hulea, M. (2009). A New Method to Obtain Non-Volatile Memory for Networks of Spiking Neurons, Memoirs of the Scientific Sections, Romanian Academy în curs de apariţie.

Loizou P.C., Dorman, M. (2006). On the Number of Channels Needed to Understand Speech, *Journal of the Acoustical Society in America.*

MAXIM Innovation Delivered™ (2006). 2.3W Stereo Speaker Amplifiers and DirectDrive Headphone Amplifiers with Automatic Level Control.

Stevens, K. (1999). *Acoustic Phonetics*, Editura Cambridge MA: The MIT Press, 204 - 214.

Teodorescu, H.N., Feraru, S.M., Trandabăţ, D., Zbancioc, M., Luca, R., Verbuţă, A., Hnatiuc, M., Ganea, R., Voroneanu, O., Pistol, L., Şcheianu, D. (2005-2007), situl Web Sunetele Limbii Române.
http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.html

Teodorescu, H.N., Trandabăţ, D., Feraru, S.M, Zbancioc, M., Luca, R. (2007). A corpus of the sounds in the Romanian spoken language for language-related education, Chapter Six, pp. 73-90. În volumul Carlos Periñán Pascual (Editor), „Revisiting Language Learning Resources", Cambridge Scholars Publishing (CSP), UK, ISBN 1-84718-156-2; ISBN 13: 9781847181565, 2007.

Teodorescu, H.N., Feraru, S.M. (2007). A study on Speech with Manifest Emotions, *10th International Conference on Text, Speech and Dialogue*, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, *Lecture Notes in Computer Science*, Springer Verlag, vol. 4629/2007, 254-262, ISBN 978-3-540-74627-0.

Wills, S. (2004). Computation with spiking neurons, *teză de dizertaţie trimisă la University of Cambridge pentru admiterea la doctorat*, 5-12.

CAPITOLUL 2

PLATFORME, DICȚIONARE ȘI CORPUSURI PENTRU PRELUCRAREA TEXTELOR

RESURSE LINGVISTICE ROMÂNEȘTI ÎN FLUX CONTINUU

DAN CRISTEA

Universitatea „Alexandru Ioan Cuza" din Iaşi

Str. Berthelot nr. 16, 700483 – Iaşi, Romania

dcristea@info.uaic.ro

Rezumat

Lucrarea prefigurează o inițiativă legislativă pentru obținerea de resurse lingvistice românești la scară mare, fără atingerea intereselor de proprietate intelectuală ori financiare ale autorilor și deținătorilor de texte. Conservarea limbii române scrise în vederea urmăririi continue a evoluției ei este perfect realizabilă din punct de vedere tehnologic. Se descrie un flux tehnologic ce are ca obiectiv colectarea tuturor textelor românești ce urmează a fi tipărite, stocarea lor pe durată indefinită și prelucrarea lor în scop științific. Sunt imaginate servicii în folosul cercetătorilor limbii cât și a proprietarilor drepturilor de autor și a autorilor de texte.

1. Introducere

Lucrarea aduce în atenție o propunere pentru colectarea, conservarea și utilizarea în scopul cercetării, a tuturor textelor publicate în România de către edituri.

Este evident faptul că, fără un efort continuu, acele limbi care, sunt cunoscute acum a avea „mai puține resurse" vor continua să fie văzute astfel chiar și atunci când, ipotetic vorbind, ar ajunge la un nivel de reprezentare în resurse echivalent nivelului actual al așa-ziselor „limbi mari". Mai mult decât atât, dacă limbile azi bogat reprezentate în resurse vor înceta acum să achiziționeze resurse, în baza raționamentului că necesitățile lor de cercetare sunt satisfăcute, în scurt timp ele își vor pierde acest avantaj. Acest lucru se datorează unui proces de deteriorare (îmbătrânire) rapidă a resursele lingvistice. Îmbătrânirea se datorează evoluției continue a limbilor dar și modificării viziunii cercetătorilor cu privire la oglindirea fenomenelor de limbă. În cazul resurselor adnotate, faptele lingvistice, care fac obiectul adnotării automate, se pot schimba de-a lungul timpului, pe măsură ce teoriile lingvistice, pe care se bazează convențiile de marcare, evoluează și pe măsură ce însăși procesele de adnotare automată se îmbunătățesc. Putem spune că nu există un final în construirea de resurse lingvistice.

În multe țări există o lege a „depozitului legal". Aceasta obligă toți furnizorii de materiale tipărite (edituri, persoane fizice sau juridice, care imprimă documente pentru public, casele de înregistrări și studiourile, Banca Națională, Monetăria Statului, Poșta Națională etc.) – să îi denumim *furnizori de resurse* – să trimită unei biblioteci naționale (care poate fi o singură unitate fizică sau un consorțiu de biblioteci), pentru conservare pe termen lung, un număr de copii pentru fiecare element imprimat destinat distribuției. Cu toate ca orizontul producțiilor de texte s-a schimbat dramatic în ultimii ani, formatul electronic surclasând formatul tipărit, după știința mea există doar încercări timide de actualizare corespunzătoare și a aspectelor juridice.

Pe măsură ce resursele lingvistice devin tot mai necesare pentru studiul limbii şi cum obţinerea lor este adesea costisitoare, problematica achiziţionării lor ar trebui să nu mai fie accidentală sau episodică, ci ar trebui să devină o politică naţională. Ceva trebuie făcut. Ar trebui să existe o lege care să protejeze resursele lingvistice ale limbilor ce se vorbesc într-o ţară şi aceasta ar trebui să fie de interes primordial. Această lucrare discută o posibilă soluţie, care, cu toate că nu este uşor de implementat, ar putea să schimbe complet scena resurselor lingvistice în viitorul apropiat.

2. Îmbunătăţirea legislaţiei cu privire la depozitul legal

O investigaţie recentă în rândul câtorva dintre cei mai importanţi producători de informaţie tipărită din Romania a scos în evidenţă faptul ca multe edituri ar fi dispuse să-şi doneze resursele în scopul cercetării. Cu toate acestea, un alt segment, care din nefericire este majoritar, nu este interesat într-o colaborare cu oamenii de ştiinţă. Aceşti producători se tem că cedarea datelor lor textuale în folosul cercetării ar echivala cu pierderea controlului asupra proprietăţii lor, ar declanşa o diminuare a profitului, sau, pur şi simplu, ignoră importanţa problemei şi nu au timp să se dedice acestor aspecte.

În realitate, nimic din toate acestea nu ar trebuie să se întâmple. Cu toate că avem nevoie de datele lor lingvistice, nu dorim ca *furnizorii de resurse* să fie păgubiţi dacă oferă textele în beneficiul ştiinţei. Ideea este de a promova o iniţiativă legislativă (să o numim **Legea depozitului electronic**) care să impună obligativitatea ca *furnizorii de resurse* să doneze datele lor lingvistice pentru supravegherea permanentă a evoluţiei limbii.

Următoarele tipuri de resurse, produse în serie, ar trebui să facă obiectul unei astfel de iniţiative legislative, indiferent dacă resursele respective sunt destinate distribuţiei comerciale sau gratuite: cărţi, broşuri, pliante, jurnale, reviste, almanahuri, calendare, partituri muzicale, materiale propagandistice cu scop politic, administrativ, cultural, artistic, ştiinţific, educaţional, religios; postere, declaraţii şi alte materiale destinate publicării în spaţii publice, teze de doctorat, cursuri universitare, documente în format electronic care conţin material lingvistic (CD-uri, DVD-uri etc.), standarde şi norme tehnice, publicaţii emise de autorităţi naţionale şi locale, colecţii de norme şi legi şi orice alt material imprimat sau multiplicat folosind metode grafice sau fizico-chimice.

La nivel practic, iniţiativa presupune existenţa unui depozit central, care este o entitate (centru de calcul, institut etc. – să-l denumim generic *Portal*), care, pe de o parte, are autoritatea legală de a primi şi stoca date oferite prin contribuţia *furnizorilor de resurse* şi, pe de altă parte, este echipat tehnic pentru a colecta, stoca pentru o perioadă de timp nelimitată şi prelucra, toate textele în format electronic ce urmează a fi tipărite şi distribuite, zilnic, într-o ţară.

Legea ar trebui să stipuleze că, prin trimiterea unei copii electronice acestui depozit naţional pentru conservarea pe termen nelimitat, nici un drept de autor sau beneficiu comercial nu va fi pierdut de către *furnizorul de resurse* sau autor. Aceasta copie va putea fi folosită, prin intermediul *Portalului*, **doar în scopul cercetării lingvistice** şi nici *Portalul* nici vreo altă persoană ori entitate nu vor putea face publice datele sau fragmente care să exemplifice o cercetare, în Internet sau într-un alt mediu, decât cu acordul proprietarului. Este clar că această lege va trebui să includă un capitol destinat

drepturilor de autor, care să fie extrem de atent conturat (COM, 2009). Alternativ, e posibil ca legea drepturilor de autor şi a proprietăţii intelectuale să necesite amendamente care să o armonizeze cu conţinutul noii legi a depozitului electronic. De asemenea, o slabă caracterizeze a măsurilor de securitate în transferul de date nu poate fi acceptabilă în formularea acestei legi.

În multe ţări se constată un interes ridicat în privinţa organizării unor servicii publice de depozitare a conţinutului electronic în biblioteci digitale[1]. Multe dintre acestea sunt arondate instituţiilor de cercetare, scopul lor fiind cu precădere acela de a păstra conţinutul ştiinţific în vederea diseminării lui cu, ori fără, restricţii. Sunt extrem de numeroase apoi iniţiativele de păstrare a cărţilor rezultate din scanări în format electronic (Galica, în Franţa[2], proiectul Guterberg[3] cu multitudinea de proiecte afiliate, serviciul de căutare oferit de Googe Books[4], proiectul Runeberg[5] pentru literatură nordică, consorţiul Open Content Alliance[6] etc.[7]).

3. Fluxul colectării datelor

Cred că o metaforă adecvată pentru *Portal* ar fi: o fabrică care procesează continuu cuvinte. Activitatea *Portalului* trebuie să înceapă cu crearea de identităţi în Portal tuturor *furnizorilor de resurse*: înscrierea la Portal a unei entităţi ca *furnizor de resurse* trebuie să se efectueze la momentul intrării în acţiune a Legii sau înainte de distribuirea primei publicaţii. În urma înscrierii, *furnizorul de resurse* va obţine un cod de identificare (*FID*) de la *Portal*. Acest cod va fi folosit pentru comunicarea cu *Portalul*, cu privire la orice publicaţie, pe toată perioada de activitate a acestuia.

Să presupunem că astăzi, *furnizorul de resurse* pregăteşte pentru publicare un nou document *D*, care a primit deja aprobarea editorială „bun de tipar". Figura 1.a prezintă fluxul de date iniţiat de acest nou element. *Furnizorul de resurse* completează un formular electronic (antet, *header – H*), care conţine informaţii de identificare a documentului, şi apoi interacţionează cu *Portalul*, încărcând *FID*-ul său, antetul *H* şi o copie editabilă a documentului *D*. *Portalul* primeşte aceste informaţii şi solicită un cod de identificare permanent (*persistent* ID – *PID*) de la o autoritate care îl poate elibera (Kunze and Rogers, 2003; Schwardmann, 2009). Atunci când primeşte un astfel de cod, stochează în depozit un pachet conţinând *PID*, *FID*, *H*, şi *D*. După aceasta, *Portalul* va întoarce *furnizorului de resurse* un mesaj de confirmare care conţine două părţi: o parte ce poate fi citită în clar şi o parte ce conţine un cod de bare. Căsuţa de confirmare trebuie să înregistreze un sigiliu al *Portalului*, împreună cu *PID*, *FID* şi *H*. Documentul, care conţine acum şi căsuţa de confirmare din partea *Portalului*, pe o manşetă, o copertă interioară sau într-o anexă, poate fi tipărit (Figura 1.b). Această căsuţă ar trebui să dovedească oricărei autorităţi însărcinate cu verificarea respectării Legii ca depozitul

[1] http://en.wikipedia.org/wiki/Digital_library
[2] http://gallica.bnf.fr/?&lang=FR
[3] http://www.gutenberg.org/wiki/Main_Page
[4] http://books.google.com/
[5] http://runeberg.org/
[6] http://www.opencontentalliance.org/
[7] O listă a proiectelor de biblioteci digitale poate fi consultată aici: http://en.wikipedia.org/wiki/List_of_digital_library_projects

electronic legal al acelui document a fost efectuat, toate informaţiile necesare pentru identificare fiind acolo.

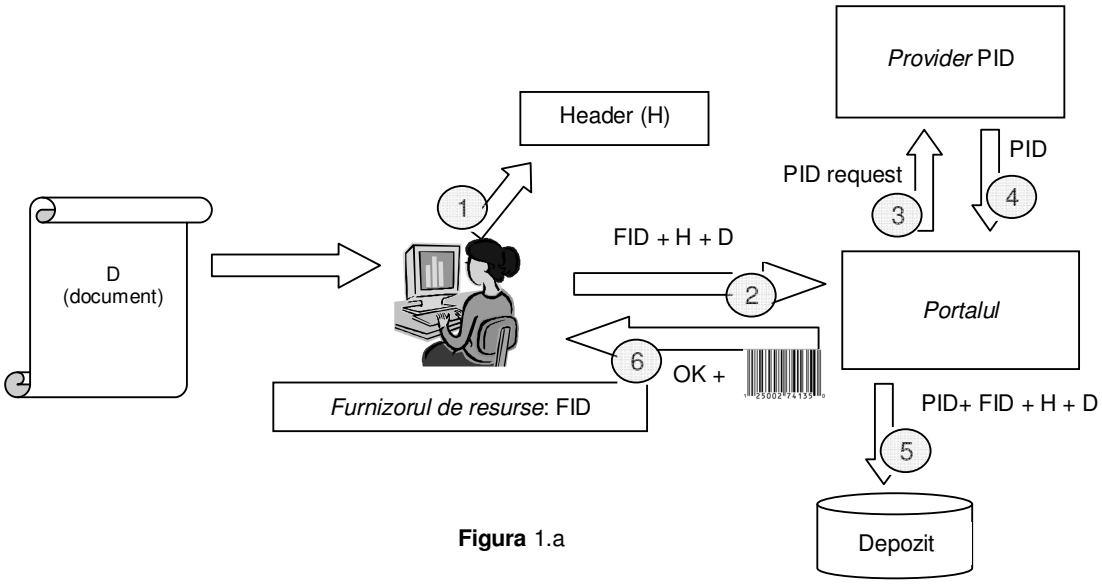


Figura 1.a

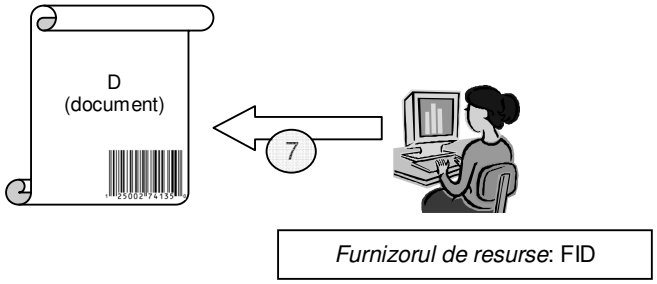


Figura 1.b

Schimbul de date detaliat mai sus între *furnizorul de resurse* şi *Portal*, incluzând şi o comunicare cu un terţ, responsabil pentru emiterea *PID*-urilor, pare greoi şi consumator de timp şi, din această cauză, total neacceptabil pentru edituri. Într-adevăr, este cunoscut faptul ca activitatea în aceste unităţi este extrem de accelerată, ele fiind constrânse să proceseze informaţiile cu mare rapiditate, adesea în condiţii de stres, cu precădere atunci când tipăresc cotidiene. În fapt însă, tot fluxul de informaţii descris mai sus poate fi automatizat, inclusiv completarea antetului *H*, care poate fi lăsată în grija unor module specializate capabile să preia elemente de identificare din documentul electronic *D*. Practic, întregul lanţ poate fi realizat la apăsarea de către editor a unui buton aflat în aplicaţia de editare. Rezultatul, materializat instantaneu, va fi caseta de confirmare, pe care editorul trebuie numai să o plaseze acolo unde doreşte pentru a fi tipărită împreună cu documentul.

4. Fluxul de prelucrare a datelor

Odată obţinute, datele din *Portal* trebuie prelucrare. În această secţiune voi descrie o listă a capacităţilor de procesare pe care *Portalul* ar trebui să le ofere.

Mai întâi, este evident că *Portalul* trebuie să posede capacităţi de stocare de foarte mari dimensiuni şi că ele trebuie special proiectate pentru conservarea datelor pe perioade de timp oricât de îndelungate. Mai apoi, el trebuie să ofere capacităţi de indexare, căutare şi acces, la diverse nivele: antet, semne lexicale (cuvinte), expresii lexicale, precum şi coocurenţe contextuale. Printre altele, acest lucru presupune ca fiecare document ce este plasat în *Portal* să fie supus unui lanţ de procesări care trebuie să includă minimal: segmentare la nivel de cuvânt, etichetare la parte de vorbire, lematizare şi indexare. Înregistrarea textului în forma iniţială este de asemenea de prevăzut, pentru că la el ar trebui să facă trimitere adnotările XML *stand-off*. Este de aşteptat ca adnotările XML şi indecşii să multiplice dimensiunea iniţială a documentelor text de câteva ori.

Pornind de la aceste funcţionalităţi de bază, o linie diferită de prelucrări se va adresa nevoilor lexicografice. *Portalul*, funcţionând, din acest punct de vedere, în acelaşi timp ca un depozit imens de resurse dar şi ca o colecţie de instrumente de prelucrare a limbii, va trebui să fie capabil să realizeze operaţiuni complexe cum ar fi: identificarea cuvintelor străine, semnalizarea unor cuvinte noi, recunoaşterea sensului cuvintelor în context, identificarea sensurilor noi, semnalizarea cuvintelor uitate (învechite), semnalizarea sensurilor care nu mai sunt folosite etc. De exemplu, se poate imagina că semnalizarea intrării în limbă a unor cuvinte noi, cât şi ieşirea unor cuvinte din circulaţie, ar trebui declanşată de existenţa, permanent actualizată, a frecvenţei de apariţie a cuvintelor pe parcursul unui interval constant de timp, atunci când această frecvenţă se plasează deasupra (ori dedesubtul) anumitor praguri, decise de o autoritate lingvistică. În mod similar, semnalizarea unui nou sens poate fi declanşată de eşecul de a cataloga sensul, desprins din context, între sensurile păstrate într-un repozitoriu de sensuri, cum ar fi, de exemplu, un dicţionar explicativ, dacă acest lucru are loc recent cu o anumită frecvenţă şi dacă modelul este suficient de stabil. Similar, sensurile uitate (învechite) sunt recunoscute atunci când regăsirea acestora scade sub un anumit prag.

Procesul care trebuie pus la baza recunoaşterii cuvintelor ori sensurilor învechite presupune monitorizarea permanentă a unui set de cuvinte luate în „colimator”. Aceste cuvinte/sensuri sunt plauzibile a fi candidaţi la dispariţie deoarece apar cu frecvenţe care scad continuu. Trebuie notat faptul că, de-a lungul unui interval de timp, criteriul frecvenţei absolute, ori chiar al uneia relative, poate să nu fie relevant, deoarece unele cuvinte sunt foarte rar folosite, cu toate că nu sunt în pericol de a fi considerate moarte (de exemplu unele neologisme ştiinţifice). Cea mai bună cale ar fi crearea câte unei fişe „personale” pentru fiecare lexical, în care să se înregistreze un set de proprietăţi (modificabile în timp), printre care şi frecventa apariţiilor de-a lungul timpului (un grafic, pe baza căruia să poată fi calculat un grad de deteriorare a utilizării), lista registrelor unde se remarcă folosirea lui (cu frecvenţele relative asociate) etc. În acest mod, problema se reduce la calcularea frecvenţei pe parcursul unui interval constant de timp, având întotdeauna ca final ziua curentă. Cineva ar putea realiza acest lucru prin simpla căutare a cuvântului în depozit şi numărarea apariţiilor care se încadrează în intervalul de referinţă – o funcţie care poate fi apelată doar o dată într-un lung interval –

să zicem de la doi la cinci ani (deoarece nu ne putem aştepta ca marcarea „învechit" să se apară brusc, de ieri până astăzi, de exemplu ...).

Este clar ca orice decizie finală cu privire la aceste poziţii trebuie să fie luată de o autoritate ştiinţifică (în speţă, Academia). Deciziile acesteia însă trebuie să se bazeze pe semnalele trimise de *Portal*, care la rândul lor trebuie să aibă la bază probe statistice clare.

Alte fluxuri de procesare pot implementa alte funcţionalităţi. Un număr de resurse în format electronic, care sunt de o importanţă majoră în menţinerea unei limbi actualizate din punct de vedere tehnologic, pot fi conectate permanent la *Portal*. Unele dintre acestea ar putea fi: Dicţionarul Tezaur oficial al limbii (eDTLR în cazul limbii române, vezi (Cristea et al., 2009)), WordNet (Fellbaum, 1998), VerbNet (Kipper et al., 2008), FrameNet (Fillmore, 1976; Atkins et al, 2003) – pentru a numi doar câteva. Presupunând ca aceste resurse devin, la un anumit moment, complete pentru limba *L*, ele trebuie actualizate permanent cu evoluţia limbii. Astfel, dinamica limbii trebuie să se reflecte şi în resurse. Dacă, aşa cum se sugera mai sus, fiecare element lexical are o fişă proprie pe *Portal*, atunci ea ar trebui să includă referinţe în toate resursele. Ca atare, fişa cuvântului *w* este legată la intrarea lui din Dicţionar, care conţine şi un inventar al sensurilor lui, iar aceste sensuri sunt sincronizate cu cele conţinute în WordNet pentru acest element lexical, precum şi cu intrarea sa din VerbNet şi FrameNet. Toate aceste resurse sunt conectate între ele şi sunt actualizate în ritm cu evoluţia limbii de către *Portal*.

Portalul poate, de asemenea, găzdui un număr de servicii adresate *furnizorilor de resurse*, cercetătorilor limbajului, consumatorilor industriali ori publicului larg. Serviciile publice ori comerciale pot fi supuse unor taxe, profitul urmând a fi întors *furnizorilor de resurse*, de exemplu proporţional cu contribuţia lunară a fiecăruia pe *Portal* (măsurată în număr de caractere) ori cu numărul de accese ale consumatorilor la propriile resurse.

Pot fi imaginate şi alte tipuri de servicii plătite, cu direcţionarea profitului către *furnizorii de resurse*. Un exemplu îl constituie postări care fac reclamă cărţilor tipărite, acces online la fragmente ale publicaţiilor ce se găsesc pe piaţă etc. Cred că posibilităţile de servicii ce ar putea fi oferite de *Portal* prin aplicarea unor rafinate tehnologii lingvistice şi pe baza cărora *furnizorii de resurse* să obţină profit sunt extrem de diverse şi ar trebui promovate cu insistenţă. Acest lucru ar putea diminua teama *furnizorilor de resurse* într-o lege care, doar aparent, afectează controlul lor asupra proprietăţii, prin aceasta, contribuind la micşorarea rezistenţei lor faţă de aplicarea ei în practică.

5. Evaluare

Este clar că o astfel de iniţiativă va aduce zilnic pe *Portal* un volum foarte mare de date lingvistice. O estimare, chiar şi grosieră, a necesităţilor de procesare şi a costurilor presupuse de un astfel de demers la nivel naţional ar trebui să ia în considerare parametri cum ar fi: numărul de edituri înregistrate, numărul mediu de publicaţii per editură per an, numărul mediu de pagini a unui obiect tipărit, numărul mediu de caractere pe pagină. Lăsând la o parte publicaţiile episodice de mici dimensiuni, ancheta

noastră cu privire la cantitatea medie de date publicate în cărţi şi periodice, într-o ţară de mărime medie din Europa cum e România, la nivelul anului 2008, a evidenţiat un volum de date în format text care nu atinge 1 GB pe zi.

Un canal cu lăţimea de bandă de 12.5 Mb/sec poate asigura cu uşurinţă transferul necesar descris în secţiunea 3, evitând blocajele chiar şi în momentele de supraaglomerare. Pentru motive de siguranţă ar trebuie asigurată o încărcare distribuită iar datele ar trebui stocate redundant (*mirroring*) în cel puţin două centre, în locaţii diferite. Cum au dovedit deja giganţii Internetului (de exemplu: Google[8]), tehnologia RAID, care grupează un număr foarte mare de mici calculatoare, este o soluţie ieftină şi care ar putea fi avută în vedere pentru a asigura conservarea pe termen lung şi pentru obţinerea unei viteze de procesare confortabilă.

6. Concluzii

Avantajele unei infrastructuri dedicate prelucrării datelor lingvistice, de o anvergură atât de mare ca cea schiţată în această lucrare, sprijinită şi de o legislaţie favorabilă, sunt greu de stabilit corect acum. În primul rând, ea oferă o soluţie completă pe termen lung pentru conservarea datelor lingvistice ale unei naţiuni, precum şi o radiografie aproape completă a evoluţiei ei diacronice. În al doilea rând, ea pune bazele unor cercetări exhaustive asupra limbii. În al treilea rând, ea aduce în atenţie o gamă largă de aplicaţii atrăgătoare din punct de vedere comercial, spre beneficiul autorilor de texte, *furnizorilor de resurse*, şi consumatorilor de limbă, prin aceasta accelerând cercetările din domeniul lingvisticii computaţionale şi al prelucrării limbajului natural şi influenţând benefic nivelul de cultură al maselor.

Succesul unei astfel de iniţiative la nivel naţional depinde foarte mult de o viziune europeană concertată. Suflul nou care se simte la momentul actual în Europa cu privire la realizarea de infrastructuri pentru procesare lingvistică, la stabilirea de standarde pentru reprezentarea datelor lingvistice, la achiziţionarea de resurse lingvistice şi susţinerea financiară comunitară a unor proiecte transnaţionale precum CLARIN[9], FlareNet[10], T4Me[11], Meta-Net[12] etc. ar trebui să creeze şi premizele unor legislaţii favorabile. Propunerea avansată în această lucrare este conformă cu alte iniţiative care încearcă să crească nivelul de conştientizare cu privire la necesitatea accesului liber la ştiinţă. În nici un caz lucrarea nu pledează împotriva proprietăţii intelectuale (Stephan, 2001), dar este în favoarea reconsiderării legislaţiei cu privire la drepturile de autor, care, în forma actuală, este extrem de restrictivă în multe cazuri relativ la utilizarea resurselor lingvistice în folosul cercetării. La urma urmei, limba noastră, aşa cum o folosim astăzi, este creaţia colectivă a naţiunii şi este supusă unor transformări (n-aş zice că acestea ar putea fi caracterizate întotdeauna drept cizelări…) permanente, realizate de către toţi vorbitorii ei. Donarea creaţiilor lingvistice pentru conservarea limbii şi cercetare, nefiind în nici un fel dăunătoare pentru creator, nici intelectual şi nici material, reprezintă, în definitiv, minimum de recompensă pe care un autor care

[8] http://infolab.stanford.edu/~backrub/google.html
[9] www.clarin.eu
[10] http://www.flarenet.eu/
[11] http://t4me.dfki.de/
[12] http://www.meta-net.eu/

utilizează limba îl datorează celor care au inventat-o, în beneficiul celor care o vor folosi în viitor.

Referințe bibliografice

Atkins, S., Rundell, M. and Sato, H. (2003). The Contribution of Framenet to Practical Lexicography, *International Journal of Lexicography*, Volume 16.3: 333-357.

COM (2009) 532 – Communication from the Commission. Copyright in the Knowledge Economy.

Cristea, D., Răschip, M., Moruz, A. (2009). Steps in Building the Electronic Version of the Thesaurus Dictionary of the Romanian Language, in Proceedings of the IVth National Conference The Academic Days of the Academy of Technical Science of Romania, ASTR – Filiala Iași și Universitatea Tehnică „Gheorghe Asachi" Iași, Ed. Agir, ISSN 2006-6586.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database, MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280: 20-32.

Kipper, K., Korhonen, A., Ryant, N., Palmer, M. (2008). A Large-scale Classification of English Verbs, *Language Resources and Evaluation Journal*, 42(1), pp. 21-40, Springer Netherland.

Kunze, J. and R.P.C.Rogers (2003). The ARK Persistent Identifier Scheme. Internet draft at http://www.cdlib.org/inside/diglib/ark/arkspec.pdf.

Schwardmann, U. (2009). PID System for eResearch. EPIC – the European Persistant Identifier Consortium, personal communication at NEERI-09, Helsinki.

Stephan, K. (2001). Against Intellectual Property. *Journal of Libertarian Studies*, 15.2, 1-53.

COMUNICAREA ELECTRONICĂ ŞI PROBLEMELE NOASTRE ORTOGRAFICE – FĂRĂ SOLUŢII?

LUCIAN CHIŞU

Muzeul Naţional al Literaturii Române, Bucureşti

lucianchisu@gmail.com

Rezumat

Acest text poate fi considerat o continuare a unei intervenţiei anterioare a autorului, apărute sub titlul Comunicarea electronică şi problemele noastre ortografice (AUSH, seria Jurnalism 2003). La acel moment, limbajul informatizat în limba română prezenta un aspect mai degrabă tranzitoriu. Autorul considera că, în scopul bunei sale utilizări, ar fi fost necesare două tipuri de intervenţii. Pe de o parte cea a specialiştilor IT în scopul realizării softului şi adaptării claviaturii la limba română, iar pe de alta, atitudinea fermă a specialiştilor (lingvişti, filologi, elite culturale) la respectarea normelor ortografice şi de limbă literară şi în cadrul limbajului informatizat, pe cale de a se generaliza în ceea ce priveşte comunicarea. Concluzia degajată din noul conţinut, incluzând perioada 2003-2010 este următoarea: deşi există soft-uri şi sunt comercializate claviaturi româneşti, deşi s-au promulgat acte normative care prevăd sancţiuni în cazul nerespectării limbii literare, abaterile, greşelile şi ezitările de la începuturi se manifestă într-o proporţie îngrijorătoare. Sunt indicate site-uri ale oficialităţilor, unele aparţinând Ministerului Culturii şi Ministerului Educaţiei, în care, departe de a fi puse în practică, reglementările ortografice sunt ocolite, în unele cazuri cu indicaţia expresă de a se evita diacriticele.

I. Începând cu ultimul deceniu al secolului trecut, utilizarea în domeniul presei a limbajului informatic şi-a făcut simţite avantajele, răsfrânte la scurt timp şi în cultura română. Pe de o parte, creşterea fără precedent a gamei şi numărului de servicii electronice (sub formă textuală, audio, grafică şi/sau video), iar pe de alta, accesul tot mai nelimitat, inclusiv cel individual şi mai ales al tinerilor, au creat premisele dezvoltării exponenţiale ale noilor forme adaptate comunicării. În primii ani ai mileniului al treilea fenomenul se generalizase, devansând tipurile de comunicare lingvistică până nu demult consacrate. Totuşi, chiar şi cei mai fideli dintre susţinătorii informatizării limbajului erau de acord cu privire la imensa cantitate de necunoscut care însoţea procesele fenomenului în discuţie[1].

[1] „În această epocă năvalnică de schimbări tehnologice rapide, toţi luptăm pentru a ne păstra propria direcţie. Progresele care se dezvăluie zilnic în comunicaţii şi tehnică de calcul pot fi înspăimântătoare şi debusolante. Este firesc, în contextul dat, să ne întrebăm dacă aceste schimbări sunt bune sau rele, dacă trebuie întâmpinate cu bucurie sau cu teamă. Răspunsul este: şi una şi alta.” (*Manifestul tehnorealist*, 2000, p. 307). Textul reprezintă rodul colaborării a 12 scriitori preocupaţi de tehnologie: David Bennahum (editor al publicaţiei „Meme”), Brooke Shelby Biggs (editorialist la „San Francisco Bay Guardian”, varianta elctronică), Paulina Borsook (scriitor), Marisa Bowe (fost redactor-şef al revistei „Word”), Simson Grafinkel (colaborator la „ Wired”, editorialist la „The Boston Globe”), Steven Johnson (scriitor, redactor-şef la „FEED”), Douglas Rushkoff (scriitor, editorialist la „Time Digital”), Andrew Shapiro (felow la Centrul Berkman pentru Internet şi Societate al Facultăţii de Drept Harvard, colaborator la „The Nation”), David Shenk (scriitor, comentator la The National Public Radio), Steve Silberman (editor în probleme de cultură la „Wired”), Mark Sthalman (scriitor, cofondator al „FEED”). Conceptul acestui document a fost schiţat de

1. Deoarece în anii imediat următori momentului utilizării computerului, în ţara noastră nu existau „tastaturi” cu claviatura adaptată la alfabetul limbii române, apăruseră adevărate modele de scriere constând în abateri de la normele ortografice. Alături de texte redactate fără semnele grafice (diacritice) ale literelor **ă**, **â**, **î**, **ţ**, **ş**, al căror înţeles deplin rezulta contextual, în tot mai extinsa comunicare pe suport electronic se dezvoltau noi tipuri de scriere în care alfabetul românesc era folosit *ad libitum*, după priceperea sau inventivitatea utilizatorilor. Chestiunea rigorii în ceea ce priveşte recuperarea sensului cuvintelor devenise pentru unii dintre utilizatori stringentă. Sensul lexical trebuia desprins din context pentru că, din cauza absenţei diacriticelor, nu rezulta clar dacă în enunţul respectiv se făcea referire la *bulgări* (de pământ ori de zăpadă) sau *bulgari*, cetăţeni ai ţării vecine („*loviti de bulgari*”), de *paturi* sau *pături* (*sinistratilor li s-au oferit paturi*), dacă în *supa de cocos*, produsul culinar provenea din carnea de *cocoş* sau din fructul exotic *cocos*, care intră ca ingredient alimentar în compoziţii asemănătoare. În pofida acestor neajunsuri, abaterile de la norme îşi croiseră calea şi sub alte forme, la fel de insolite[2]. Metodele de aplicaţie fiind iluzorii, ele înseşi constituiau erori sau devieri de la normă şi ilustrau totodată accesul dificil, greoi, complicat şi inutil la un sistem (tastatură) impropriu.

Lucrul se datora în primul rând unei carenţe legislative, dar, în ansamblu, fenomenul prezenta o dublă cauză.

a. Cea dintâi, de natură tehnică, necesita realizarea softului şi a tastaturilor pentru limba română, fiindcă iniţial cele comercializate în România fuseseră proiectate pentru uzul în limba engleză. Sarcina realizării „programelor” pentru diacritice şi a tastelor aferente alfabetului românesc cădeau exclusiv în seama programatorilor IT şi inginerilor de sistem din ţara noastră. Chestiunea a fost întrutotul depăşită şi a devenit de domeniul trecutului.

b. Consecinţă nemijlocită a celei dintâi, a doua cauză servea pragmatismului greşit înţeles: utilizarea corectă a limbii române în scrierea electronică trecuse pe un al doilea plan, cedând întâietate beneficiilor, fapt care a favorizat instalarea aproape nestânjenită a unui autentic haos ortografic. Contau mai mult imensele avantaje ale scrierii electronice constând în fiabilitate, actualizare, conservare, multiplicare, dar mai ales eficienţa, rapiditatea şi confortul comunicării electronice, spaţiul şi timpul fiind contrase în ceea ce numim oarecum impropriu „mediul virtual”. Cu alte cuvinte, comunicarea electronică se afla în coliziune cu alfabetul şi ortografia noastră normativă, situându-se, spre nemulţumirea unora dintre utilizatori, într-un punct tranzitoriu de care nimeni nu părea a fi responsabil să ia atitudine. S-a instalat, nu fără temei, teoria formelor fără fond, valabilă în tipuri de situaţii în care „forţele reacţionare”, înşelate în aşteptările lor, declanşează conflictul dintre procesul vieţii (care nu ştie ce-i răgazul) şi formele de exprimare culturală (care nu ştiu ce-i schimbarea). Sentimentul nemulţumirii se conturează în astfel de situaţii sub forma crizei.

Shapiro, Shenk şi Johnson. A fost publicat pe 12 martie 1998. Poate fi citit în versiune originală la *www. technorealism.org.* Vezi, Ana Tăbîrcă, traducătoarea *Manifestului tehnorealist* (***, 2000).

[2] Exemple privind improprietăţile semantice ale textelor scrise fără diacritice, cu mult mai numeroase, care proliferaseră în limbajul informatic al anilor 2000, am oferit în (Chişu, 2002) şi în studiul din (Chişu, 2003).

După trecerea unui număr de ani, dezavantajele aspectului prim (tehnic) au fost anulate, însă consecinţele începuturilor n-au dispărut, putându-se afirma că „problema", deşi părea a fi fost rezolvată, încă persistă.

În acest context ambiguu, singura certitudine o reprezintă generalizarea scrierii pe suport electronic, care se manifestă ca un fenomen (inter)naţional tocmai ca urmare a imenselor foloase pe care le generează. Ca pretutindeni pe glob, instituţiile oficiale şi private din România, precum şi uriaşa masă de utilizatori individuali au trecut la noul format. Indiferent de caracterul oficial sau particular al textelor, majoritatea covârşitoare a lor este procesată electronic şi abia după aceea transferată pe suportul tradiţional (hârtia). Utilizăm fără îndoială, la scară naţională, limbajul informatic cu extraordinarele sale disponibilităţi şi, în continuare, însoţite de „reflexele" vechilor abilităţi.

2. Este, iarăşi, o certitudine că, în privinţa limbilor „culte", limbajul informatic reprezintă tehnic cea mai spectaculoasă ascensiune. Dar, în ceea ce priveşte culturile, ale căror sisteme lingvistice includ în evoluţia lor alte semne grafice decât cele consacrate pentru limba engleză, limbajul informatic a exercitat, cel puţin pentru o perioadă de timp, un altfel de impact. Limbajul constituie fundamentul comunicării între oameni şi în mod indiscutabil pentru toţi aceştia poartă profunde conotaţii emoţionale şi culturale provenind din moştenirea lor literară, istorică, filosofică şi educaţională. De aceea, limba maternă nu trebuie să reprezinte un obstacol în calea accesului la cunoaşterea multiculturală umană disponibilă în noul mediu. Dacă utilizarea ei ortografică devine, cel puţin pentru unele semne grafice, aleatorie, obstacolele nu întârzie să apară.

II. Cu toate că, pe fond, chestiunea utilizării limbajului informatic în spaţiul cultural românesc şi-a găsit soluţia optimă, între altele şi ca urmare a emiterii şi promulgării unor legi (Hotărâri de Guvern), numeroşi dintre utilizatorii săi împărtăşesc sentimentul că de la soluţie la rezolvarea celui de-al doilea aspect, cel practic, rezistenţa în faţa regulilor, normelor şi corectitudinii continuă să se manifeste în proporţii de masă, chiar dacă efectele recuperatorii îşi vor face simţite în cele din urmă, cândva, consecinţele. Aceştia nu caută vinovaţi sau vinovăţii dar, pe de altă parte, nici nu pot accepta că s-a asistat (pasiv) la un fenomen… natural, deci greu de stăpânit, incontrolabil. În toţi aceşti ani de evoluţie nestânjenită a unui fenomen care devenea pe zi ce trece previzibil şi ireversibil o nouă formă generalizată de comunicare, ar fi fost de aşteptat reacţii şi atitudini din rândul specialiştilor limbii. Punctul de vedere al acestora asupra respectării normelor ortografice şi, implicit al limbii noastre literare în comunicarea electronică ar fi avut cea mai mare greutate şi autoritate. În acest interval, reacţia dumnealor a fost foarte puţin sau deloc vizibilă, cei mai mulţi resemnându-se a consemna adevărul că limba română se află într-o grea suferinţă. Totuşi, domniilor lor, doctorilor în… litere le-ar fi revenit un rol mai activ.

1. Era, de asemenea, datoria tuturor intelectualilor, de toate formaţiile umaniste ori scientiste, să ia act de autentica mutaţie intervenită, care favorizează saltul în SI-SC şi în viitorul comun. Această formă de sprijin ar fi constituit cel mai important element de solidarizare cu specialiştii în limbaje (lingvişti, filologi) care, altfel, împart o soartă ingrată, ştiut fiind că primii dintre aceştia studiază mai întâi evoluţiile şi tendinţele limbajului, mereu schimbător, spre a formula eventual recomandări şi finalmente

norme, respectate cu sfinţenie de filologi şi nu numai de ei, ci de toţi indivizii care formează societatea.

a. Trebuie ţinut cont, de asemenea, că promovarea limbii române în SI-SC sub aspectele sale scris-vorbit a devenit, cu un cuvânt ale cărui semnificaţii sunt mai degrabă arhaice, apanajul altor domenii de investigare, în sensul acesta vorbindu-se curent despre „ingineria limbajului", aplicaţie care vine să sublinieze aspectele legate de validarea experimentală a ipotezelor ştiinţifice prin modele riguros verificate. În format electronic, limbajul scris – vorbit ocupă numai o avanscenă (de resursă lingvistică), domeniul de cercetare devenind prelucrarea automată a limbajului natural. Acesta din urmă reflectă progresele ştiinţifice şi tehnologice care-l desprind de momentul iniţial (lingvistica formală) spre a accede în zonele generării de modele abstracte (lingvistica matematică) şi spre lingvistica computaţională, constând în prelucrarea limbajului natural ca *summum* al teoriilor şi modelelor de inteligenţă artificială. Aceste forme de cunoaştere a limbajului abordează problematica de comunicare om-calculator.

b. Este, de asemenea, demn de consemnat faptul că, încă din primii ani de manifestare a limbajului informatic, specialiştii noilor arii de cercetare, au ţinut să accentueze ideea că dezvoltarea armonioasă a societăţii informaţionale bazată pe cunoştinţe este posibilă doar prin promovarea informaţiei şi accesului cu caracter multilingv şi multicultural. Mai mult chiar, în raportul special al comisiei Europene (*Towards a European Language Infrastructure*), raportorul A. Danzin atrăgea atenţia asupra următorului pericol: „În era electronică, este esenţial pentru supravieţuirea unei limbi ca ea să fie folosită în sistemul de informare electronică", rezultând cel puţin indirect că procesul de informatizare a unei limbi naturale nu înseamnă în niciun chip stâlcirea ei".

Sub aceste aspecte, datoria celor mai responsabili dintre utilizatorii limbii noastre rămăsese doar un deziderat.

III. La cele afirmate până acum este necesar să adăugăm o serie de aspecte particulare ce se regăsesc în evoluţia istorică a comunicării scrise în limba română.

1. Înseşi „realităţile" interne, istorice, ale scrierii în limba română, au cunoscut etape punctate de continue dispute şi neînţelegeri de-a lungul epocilor. În ceea ce priveşte „biruinţa scrisului în limba română", înaintaşii noştri s-au servit iniţial de un alfabet cu grafie slavonă de 43 de slove (alfabetul chirilic). Concomitent cu utilizarea greoiului alfabet chirilic, întrupând „sufletul românesc în veşmânt slavon" (cum l-a numit Nicolae Cartojan), la sfârşitul secolului al XVIII-lea, învăţaţii ardeleni se foloseau de grafia latină[3], dar intenţia reprezentanţilor *Şcolii ardelene* se îndrepta spre utilizarea unui alfabet etimologic, care ar fi dus la o supărătoare ruptură între modul de a scrie şi pronunţa cuvintele. Astfel de fapte erau şi mai mult adâncite de caracterul incipient, pentru acele vremuri, al utilizării grafiei. „Regulile ortografice propuse în urma discuţiilor derulate între 1866 şi 1881 de Academia Română întâmpinau dificultăţi deosebit de mari în răspândirea lor şi rezistenţă pe măsura dificultăţii lor" (Zugun, 1992).

[3] Se ştie, spre exemplu, că Ion Budai-Deleanu a optat primul pentru grafiile ***ţ*** şi ***ş***, aşa cum sunt redate astăzi.

Considerând că au fost schiţate câteva elemente relevante pentru începuturile alfabetului şi evoluţiei ortografiei la noi, informaţii care se regăsesc în studii şi cărţi publicate de numeroşi specialişti (Avram, 1990), (Beldescu, 1985), (Hristea, 1981), (Nica şi Cureteanu, 1980), (Iordan, 1978), (Şuteu, 1976), (Graur, 1974), (Ciobanu şi Sfârlea 1970), (Rosetti, 1967), consemnăm că discuţii polemice, urmate de propuneri, reforme şi hotărâri ortografice au avut loc în anii 1881, 1904, 1925, 1932, 1953, 1957 şi 1991, acestea nestingându-se[4] niciodată.

Reţinem însă, că în atmosfera tensionată, dacă nu uneori incendiară, a acestor dispute, totuşi s-a convenit asupra necesităţii unui mai pronunţat caracter coercitiv al normelor ortografice, după cum rezultă şi din unele nu foarte îndepărtate în timp luări de poziţie în istoricul problemei.

2. La avatarurile alfabetului şi normelor ortografice succint enunţate mai înainte, se adaugă hotărârea academică ratificată în şedinţa Adunării Generale a Academiei Române din 31 ianuarie 1991, a cărei contestare a avut drept efect scindarea alfabetului. Numeroase edituri şi publicaţii, unele de prestigiu, au ignorat hotărârea în discuţie[5].

În plus, concomitent cu „schisma alfabetică" din 1991, odată cu pătrunderea masivă în societatea românească a tehnologiilor multimedia aveau să se ivească noi situaţii conflictuale. În uz s-a instalat şi cea de a treia formă, scrierea electronică fără diacritice, cu variantele ei opţionale. Ideea eludării convenţiei şi ignorarea efectului normativ a generat frecvent puncte de vedere orgolioase, unul dintre acestea fiind exprimat de un cunoscut critic literar, universitar şi academician, director al celei mai cunoscute publicaţii literare româneşti: „Iar dacă dorim cu tot dinadinsul să facem convenţia cât mai firesc-utilă, atunci, date fiind schimbările produse în scriere de calculatoarele moderne [noi am fi spus *de utilizatorii calculatoarelor*], ar fi poate necesar să abandonăm sedilele de la ***ţ*** şi ***ş*** şi să scriem ***tz*** şi ***sh***, [internaţionalizând] cumva scrierea, sau chiar să scriem, ca portughezii, pe ***ă*** ca ***a***, indiferent de pronunţie, şi pe ***î*** ca ***i*** (cum de altfel calculatorul o şi face!) [*noi am sublinia că nu calculatorul, ci utilizatorii o ...fac!*]. Dar oare reformiştii noştri vor fi de acord? Nu cred, fiindcă problema lor nu e de a avea o convenţie care să fie lesne respectată, ci, una, cum să zic? *respectabilă*"

[4] În 1881, prin contribuţia decisivă a junimiştilor, reveniţi la Academie după slăbirea inevitabilă a poziţiei, dominante până atunci, a latiniştilor, a fost salutară impunerea ortoepiei limbii, prin influenţa scrierii fonetice asupra vorbirii, putându-se crea, numai astfel, o benefică circularitate sistemică *vorbire îngrijită — scriere fonetică — ortoepie*, imposibilă în cazul unei ortografii bazate pe etimologism. Între 1881- 1904 s-a realizat trecerea de la etimologism la fonetism, etapă de tranziţie ce „ a întărit unele modalităţi de scriere care s-au impus, ca, de exemplu, scrierea vocalei [**ă**] prin a cu un semn diacritic deasupra, a consoanelor [**ş**] şi [**ţ**], prin s şi respectiv t, cu sedile ş.a.m.d." (apud Petru Zugun, 1990, p. 42). Un alt proiect academic a fost redactat în 1932 de Ov. Densusianu. El a avut drept moment iniţial discuţiile începute încă din 1925 de Sextil Puşcariu, reprezentând modificări sau precizări aduse reformei ortografice din 1904. O nouă simplificare şi clarificare a ortografiei româneşti realizează Academia în 1953 cu importanta completare din 1965, când este normată, tot de Academie, folosirea literei ***â*** în familia cuvântului *român*.
În anul 1991 Academia Română, prin preşedintele ei de atunci, dl. acad. Mihai Drăgănescu, a luat hotărârea de a recomanda câteva amendamente la normele ortografice ale limbii române din 1953: a) utilizarea lui ***â*** în locul literei ***î*** în interiorul cuvintelor; păstrarea literei ***î*** în interiorul cuvintelor dacă acestea sunt rezultatul unui cuvânt compus, în care cel de-al doilea element lexical îl conţine pe ***î*** iniţial; b) utilizarea formei ***sunt*** în locul formei ***sînt***, procedându-se în mod similar cu formele flexionare. Această hotărâre ratificată în şedinţa Adunării Generale a Academiei Române din 31 ianuarie 1991, a stârnit vii proteste şi noi puncte de vedere polemice. În aceeaşi perioadă apare şi se dezvoltă exponenţial scrierea electronică, mijlocită de computer.

[5] Astăzi avem situaţia, cel puţin bizară, a editurilor care, atunci când tipăresc manuale, şcolare folosesc ortografia academică, iar când editează alte cărţi o utilizează pe aceea ...contestatară.

(Manolescu, 2002). Cei de la care era de aşteptat „tăierea" acestui veritabil nod gordian nu s-au pronunţat, dar acest derapaj n-a fost deloc pe placul informaticienilor.[6]

3. Iată cum se (re)prezintă limba română în format electronic pe unele dintre site-urile oficiale, la ora de faţă:

a. CNCSIS (Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior), organ consultativ al Ministrului Educaţiei, Cercetării, Tineretului şi Sportului, exprimând totodată punctul de vedere al comunităţii ştiinţifice în ceea ce priveşte politica cercetării ştiinţifice apare postat fără diacritice. Unele fişiere sunt dedicate consultării în limba română, altele în limba engleză, dar între acestea nu există practic nicio diferenţă. La rubrica *Instrucţiunile privind completarea formularului „cerere de finanţare"* se face precizarea: „este OBLIGATORIU ca în timpul completării formularului să nu introduceţi caracterul ghilimele, caractere diacritice". Tot astfel stau lucrurile, adică fără prezenţa diacriticelor, şi cu CNFIS (Consiliul Naţional pentru Finanţarea Învăţământului Superior), UEFISCSU (Unitatea Executivă pentru finanţarea Învăţământului Superior şi a Cercetării Ştiinţifice Universitare), parţial cu AFCN (Administraţia Fondului Cultural Naţional), care se subordonează Ministerului Culturii şi Patrimoniului Naţional, unde, în cadrul rubricii „întrebări frecvente", se recomandă evitarea folosirii normelor obligatorii de scriere, cu diacritice.

b. Notariatele folosesc în documentele legalizate diacriticele numai la cererea expresă a clientului. În celelalte situaţii, formula fără diacritice este generalizată, în pofida faptului că pot apărea confuzii supărătoare, mai ales de nume.

c. În marea majoritate a Universităţilor din ţara noastră sunt acceptate Teze de licenţă şi de masterat redactate în mod frecvent fără diacritice. S-au întâlnit cazuri, însoţite de naive surprinderi, că ar fi greşită redactarea fără diacritice, inclusiv la limba şi literatura română. Faptul nu este în măsură să enerveze, ci, dimpotrivă, să se generalizeze.

d. Cele mai multe dintre site-urile de pe net perpetuează fără jenă situaţia descrisă mai sus.

Exemplele ar putea continua dar, din cele prezentate, se detaşează următoarea observaţie: raportul dintre scrierea în limba română literară şi celelalte formate înclină majoritar în favoarea celui nerecomandabil, greşit. Singura concluzie, din păcate „viabilă", se referă la adevărul că generalizarea scrierii electronice a devenit o realitate de necontestat.

4. Dacă nu vom trece sub tăcere că acţiunea globalizării are consecinţe devastatoare asupra culturilor identitare, atunci trebuie să invocăm şi studiul lui (Bârlea, 2009) care demonstrează indubitabil, prin intermediul unor cercetări efectuate de specialişti ai Comunităţii Europene, că dispariţia repercusiunile asupra limbilor

[6] „Există o sumedenie de situri unde limba română apare trunchiat, adică extirpată de diacritice. Prima explicaţie care ţi se serveşte este problema anumitor greutăţi dificil de contorizat. Dacă acum ceva timp exista un sâmbure de adevăr, acum e foarte greu de susţinut o astfel de explicaţie, nocivă după părerea mea. Cu mult mai grav e că există profesori în învăţământul românesc care în loc să promoveze corecta folosire bat apa-n piuă cu subterfugii. Dar nu numai aici vezi astfel de «scăpări». Există o serie de instituţii culturale, edituri, reviste de exemplu, unde o astfel de explicaţie îşi face loc cu coatele. Am fost surprins să aud din gura unui critic literar şi conducător de revistă o astfel de inepţie. La contra argumentele mele, replica a fost un soi de făcut cu mâna din categoria «merge şi aşa»" (Dan Iancu, redactor şef „PCMagazine România", p. 4.).

naţionale reprezintă un fenomen real şi destul de ameninţător, ca şi topirea calotei glaciare. Printre cauzele dispariţiei limbilor, autorul articolului citat enumeră: reacţia tinerelor generaţii faţă de limba maternă, potenţialul social-economic al comunităţilor de vorbitori, relaţia scris-oral, politicile oficiale. Desigur, faptele prezentate mai înainte nu reprezintă, în acest moment, un pericol real pentru limba română. Din studiul cercetătorului înainte amintit, rezultă, pe bază de date şi informaţii, că la fiecare două săptămâni dispare câte o limbă având mai puţin de 200.000 de vorbitori. În acest ritm, accelerat de noile media, ştergerea identităţilor se va produce în următorii o sută de ani. E bine? E rău? Apariţia unui *model unic*, globalizant, pentru noile situaţii create, generează întrebări şi solicită răspunsuri la interogaţii diverse. Nu se schimbă, prin urmare, funcţiile sau domeniul comunicării prin limbaj, ci, prin noile tehnologii instalate, pe lângă imense avantaje, se profilează şi un nou mediu şi tip de cultură, aflat în conflict nedeclarat cu vechile modalităţi de comunicare tradiţională, responsabile cu metabolismul organismelor culturale. E bine? E rău?

Ca instrument (proteză) a informaţiei şi cunoaşterii, în pofida faptului că a pătruns adânc în evoluţia societăţii, în ţara noastră nu a existat şi nu există o politică de stat coerentă cu privire la această mare invenţie, cea mai discutabilă dintre problemele ridicate în discuţie fiind aceea a patrimoniului naţional care deţine un specific naţional sau identitar puternic conturat.

IV. Concludem, aşadar, că, privitor la dezbaterile asupra scrierii, inclusiv ale celei pe suport electronic, în lumea filologilor formula *grammatici certant*, îşi pune amprenta, etern actuală. Prin urmare, în absenţa unor decizii ferme, deşi instituţionalizate, s-a ajuns la situaţii realmente de neînţeles.

Programul conferinţei *ConsILR 2010* (*Resurse lingvistice şi instrumente de lucru pentru prelucrarea limbii române*), în cadrul căruia îşi unesc forţele informaticieni, cercetătorii în inteligenţa artificială şi lingvişti se situează la polul diametral opus. Chestiunile supuse dezbaterii privesc *realizarea de resurse lingvistice româneşti*, *textuale ori vorbite, în forma originară ori adnotată, crearea de corpusuri româneşti reprezentative, realizarea de colecţii lexicografice româneşti în format electronic, tehnologii lingvistice aplicate limbii române, aplicaţii în care au fost utilizate tehnologii lingvistice pentru limba română, realizări de lingvistică teoretică cu aplicaţii în tehnologia limbii române, proiecte de cercetare ce implică dezvoltarea de resurse şi instrumente dedicate limbii române*.

Faţă de cele prezentate în această intervenţie, *Conferinţa* găzduită de Muzeul Naţional al Literaturii Române devine un spaţiu aulic, locul perfect al schimbului de idei pentru viitorul Si-SC. Şi, totuşi, rămâne sentimentul extrem de disconfortant că între realitatea palpabilă a utilizării limbii române în comunicarea de tip electronic şi obiectivele *Conferinţei*, repet de o utilitate maximă, se interpune, metaforic vorbind, povestea episodului căderii oraşului Constantinopol, când, în timp ce fortăreaţa era escaladată de păgâni, în cercurile înaltelor spirite (religioase) din acel loc se purta o discuţie extrem de animată despre ...sexul îngerilor. Desigur, afirmaţia trebuie luată numai ca o metaforă, poate nereuşită, dar căreia am încercat să-i aducem câteva argumente, după opinia noastră, indubitabile.

Tocmai de aceea, se poate admite că, paralel cu aceste inițiative atât de demne de interes, aspectele „practice” ale uzului limbajului în format electronic, semnalate pe parcursul acestei intervenții, merită readuse în actualitate, în fața participanților la *ConsILR 2010*.

Referințe bibliografice

*** (2000), Manifestul tehnorealist, în „Secolul XX” , nr. 4-9, 2000 (421-426), pp. 307-311 (traducerea Ana Tăbîrcă).

Avram, M. (1990), Ortografie pentru toți (30 de dificultăți), Editura Academiei Române, București.

Bârlea, P. G. (2009), Dispariția limbilor - catastrofă umană sau formă naturală de schimbare lingvistică?, în „Caiete critice”, nr. 8-9 (262-263)/2009, pp. 97-117.

Beldescu, G. (1985), Ortografia actuală a limbii române, Editura Științifică și Enciclopedică, București.

Chișu, L. (2002), O chestiune urgentă, în „Tribuna învățământului”, an LIII, nr. 634 (2515), 18-24 martie 2002, p. 1; 2.

Chișu, L. (2003), Comunicarea electronică și problemele ortografiei noastre, în „Analele USH”, Seria Jurnalism, IV, nr.4, pp. 31-48.

Ciobanu, F., Sfîrlea, L. (1970), Cum scriem, cum pronunțăm corect (norme și exerciții), Editura Științifică, București.

Graur, Al. (1974), Mic tratat de ortografie, Editura Științifică, București.

Hristea, T. (1981), Sinteze de limba română, ediția a II-a, revăzută și mult adăugită, Editura Didactică și Pedagogică, București.

Iordan, I. (1978), Istoria lingvisticii românești, Editura Științifică și Enciclopedică, București.

Manolescu, N. (2002), Cum scriem, în „România literară, An XXXV, nr. 38, 18-24 septembrie 2002, p. 1.

Nica, M., Cureteanu, S. (1980), Predarea ortografiei în gimnaziu, Editura Didactică și Pedagogică, București.

Rosetti, Al. (1967), Introducere în fonetică, Editura Științifică, București.

Șuteu, F. (1976), Influența ortografiei asupra pronunțării literare românești, Editura Academiei, București.

Zugun, P. (1992), Ortografia limbii române. Trecut, prezent, viitor, Institutul European, Iași, 1992, p. 40.

FACTORII CARE INFLUENȚEAZĂ MOMENTUL OPTIM DE MIGRARE LA DIACRITICELE CORECTE ÎN LIMBA ROMÂNĂ[1]

BOGDAN STĂNCESCU

S.C. Moongate Video Production srl, București – România

bogdan@moongate.ro

Rezumat

Preconizez că în perioada 2010-2015 va avea loc în cea mai mare parte migrarea conținutului de limbă română de la utilizarea caracterelor cu diacritice cu sedilă (conform ISO-8859-2) la caracterele cu diacritice corecte cu virgulă (conform Unicode 3.0). Acest document încearcă o sinteză a factorilor care influențează, de la caz la caz, momentul ideal de migrare.

1. Introducere

Semnele diacritice din partea de jos a caracterelor românești „ş" și „ţ" sunt virgule. Acest detaliu este de la sine înțeles pentru orice vorbitor nativ de limbă română: este un fapt nedisputat care se învață în clasele primare și nu mai trebuie repetat niciodată în mod explicit. Iar asta într-o asemenea măsură încât însăși Academia Română nu a simțit nevoia să se pronunțe în această privință decât în anul 2003, și chiar și atunci numai pentru că a răspuns unei întrebări explicite în acest sens.[2]

Atunci când a fost creată prima codare a caracterelor pentru Europa de Est, în 1987 (Latin-2[3]), caracterele pentru limba română au fost comasate cu cele pentru alte limbi din această zonă geografică scrise în mod uzual cu grafie latină, precum ceha, maghiara, poloneza și altele. Între limbile asociate acestui standard, limba română este singura care folosește caracterele „ş" și „ţ", *sau orice caractere similare din punct de vedere vizual*.

Caracterul „ş" din limba română („*s cu virgulă*") este foarte similar din punct de vedere vizual cu litera „ş" din limba turcă („*s cu sedilă*"). Diferența grafică dintre cele două semne diacritice este aproape insesizabilă pentru mărimi mici de text[4] (vezi Figura 1).

Standardul Latin-2 nu a fost niciodată asociat limbii turce[5]. Cu toate acestea, caracterele „ş" și „ţ" au fost definite în acest standard, pentru limba română, drept „*s cu sedilă*" (caracterul turcesc), respectiv „*t cu sedilă*" (caracter practic nefolosit în nici o limbă).[6]

---

[1] 9 MAI 2010, VERSIUNEA 2.0 (Cea mai actualizată versiune a acestui document, împreună cu alte resurse conexe, se vor găsi întotdeauna la adresa http://www.moongate.ro/products/diacritice/)

[2] Vezi http://www.secarica.ro/html/s-uri_si_t-uri.html.

[3] Standardul ISO/IEC 8859-2, cunoscut și ca Latin-2.

[4] Vezi și nota .

[5] Caracterele pentru limba turcă au fost incluse inițial în standardul ISO/IEC 8859-3 (Latin-3), creat pentru Europa de Sud; câțiva ani mai târziu a fost creat un standard special pentru limba turcă, ISO/IEC 8859-9 (Latin-5).

[6] Această discrepanță dintre uzanța autohtonă și standardele ISO a fost probabil cauzată de faptul că înseși documentele românești care descriau semnele diacritice respective le-au numit sedile timp de aproape două secole – vezi http://www.capisci.ro/articole/Sedile_%C3%AEn_rom%C3%A2n%C4%83

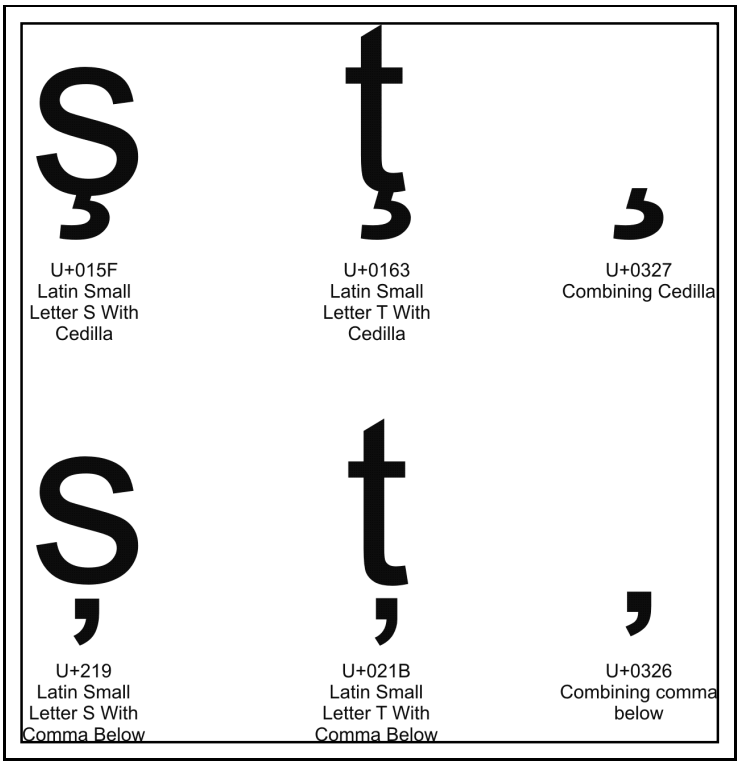


Figura 1: Caracterele în cauză, mărite pentru vizibilitatea semnelor diacritice.
Primul rând conține caracterele „*s cu sedilă*”, „*t cu sedilă*” și semnul diacritic sedilă.
Al doilea rând conține caracterele „*s cu virgulă*”, „*t cu virgulă*” și semnul diacritic virgulă.

În anul 1997 compania Apple a schimbat caracterele din standardul proprietar MacOS Romanian în așa fel încât să utilizeze semnele diacritice corecte pentru „ş” și „ţ”.[7] În același an Asociația de Standardizare din România a protestat pe lângă ISO în privința standardului Latin-2, însă singura modificare făcută un an mai târziu de către ISO a fost adăugarea unei note care permitea interpretarea semnelor respective drept „*s cu virgulă*”, respectiv „*t cu virgulă*”, iar asta numai în măsura în care expeditorul și destinatarul mesajului se puneau cumva de acord în această privință.[8]

În anul 1999 a apărut prima versiune a standardului Unicode care să conțină caracterele corecte românești „*s cu virgulă*” și „*t cu virgulă*”. Dificultățile de adoptare a standardului Unicode în formă completă au cauzat însă întârzieri de mai bine de zece ani în adoptarea sa pe scară largă.[9]

---

[7] http://unicode.org/Public/MAPPINGS/VENDORS/APPLE/ROMANIAN.TXT

[8] „*Note - Subject to the agreement of originator and receiver, in information interchange the letters S and T WITH CEDILLA BELOW may be used to substitute for the letters S and T WITH COMMA BELOW*”, http://www.secarica.ro/html/s-uri_si_t-uri.html

[9] Caracterele cu semne diacritice corecte sunt incluse și în standardul ISO/IEC 8859-16 (Latin-10), publicat în 2001. Latin-10 este suportat în navigatoarele de Internet Firefox, Chrome și Opera și în sistemele de operare Mac OS, de la versiunea 10.4 (http://www.opensource.apple.com/source/CF/CF-368.25/String.subproj/CFStringEncodingExt.h), și Linux (http://www.kernel.org/doc/man-pages/online/pages/man7/latin10.7.html). Deși prezintă unele avantaje în fața Unicode (în special pentru că standardele ASCII extinse sunt single-byte), adoptarea pe scară largă a Unicode a făcut

Așa se face că timp de douăzeci de ani aproape toate textele scrise în limba română în medii informatice au fost scrise fie fără semne diacritice, fie cu semne diacritice greșite. Abia sistemul de operare Windows Vista, apărut la sfârșitul anului 2006, a fost primul sistem de operare utilizat pe scară largă de consumatorii de conținut care să folosească în mod nativ caracterele corecte pentru limba română. Chiar și așa, penetrarea lentă a acestui sistem de operare și a celor ulterioare face ca migrarea către utilizarea în practică a semnelor diacritice corecte să fie încă și astăzi, în 2010, mai mult un subiect de discuție sporadică decât obiectul unor acțiuni concrete.

2. Contextul de aplicabilitate al acestei lucrări

În această lucrare voi utiliza termenul de *colecție de text* în sens cât se poate de abstract și de cuprinzător: orice colecție de texte în limba română stocate electronic, indiferent de formatul concret de stocare sau de modul de prezentare. De la jurnal, revistă, carte sau enciclopedie tipărită până la mesaje e-mail, site-uri Internet sau etichete de text din cadrul aplicațiilor software, toate vor migra mai devreme sau mai târziu de la caractere cu sedile la caractere cu virgule.[10]

În ceea ce privește *factorii abstracți* care influențează alegerea momentului optim de migrare, am căutat să identific o structură suficient de generică încât să fie aplicabilă oricărei situații practice, în contextul definiției cuprinzătoare din paragraful anterior.

Pe de altă parte, *datele statistice concrete* prezentate în această lucrare sunt specifice numai colecțiilor de text care satisfac simultan următoarele criterii independente:

1. sunt consultate prin intermediul unui navigator de Internet (*web browser*);
2. sunt consultate de consumatori eterogeni în privința platformei software utilizate.

Dacă măcar unul dintre criteriile de deasupra nu se aplică, atunci trebuie *ignorate complet toate datele statistice concrete* din această versiune a acestui document.[11] În acest caz trebuie utilizate resursele indicate la sfârșitul acestui document (dacă se aplică), sau trebuie căutate și adaptate datele statistice concrete asociate situației concrete la structura prezentată aici.

În privința dimensiunii temporale a deciziei de migrare am decis să nu includ niciun fel de date concrete, întrucât nivelul estimat de eroare al oricărei predicții de această natură ar fi prea mare pentru orice scop practic. Am ales în schimb să actualizez acest document și resursele conexe pe măsură ce evoluează situația (vezi nota **Error! Reference source not found.** sau ultimul paragraf din această lucrare).

---

ca, pentru moment cel puțin, Latin-10 să nu fie adoptat pe scară largă; un alt factor decisiv în această privință este și faptul că sistemele de operare și navigatoarele de la Microsoft nu suportă acest standard..

[10] Eu personal am întâlnit această problemă în contextul discuțiilor de la Wikipedia în limba română; acolo mi-am și format și sintetizat în mare parte argumentele expuse în această lucrare, interacționând cu comunitatea de voluntari din cadrul proiectului respectiv.

[11] De exemplu dacă (1) este vorba despre o aplicație client-side, (2) este o aplicație online dedicată clienților care folosesc terminale mobile, (2) este o aplicație (online sau offline) care rulează numai pe o platformă anume, sau (1+2) este o aplicație client-side pentru terminale mobile.

3. Alegerea momentului: de ce este important

După cum am arătat mai sus, diferența grafică dintre cele două variante de caractere este în cea mai mare parte a timpului nesemnificativă.[12] Din acest motiv *nu există nicio presiune naturală considerabilă pentru adoptarea semnelor corecte* – practic toți consumatorii de conținut pot interpreta corect semnele diacritice „vechi", iar majoritatea acestora nu sunt oricum la curent cu această problemă grafică minoră. Chiar și într-un sens mai profund chestiunea este la fel de neimportantă: *într-un text scris în limba română distincția dintre cele două tipuri de semne diacritice nu are valoare semantică*, deoarece utilizarea uneia dintre variante în defavoarea celeilalte nu aduce niciun plus de informație, indiferent de felul în care este interpretat textul.[13]

Prin urmare avem de-a face cu o *problemă semnificativă de interoperabilitate cauzată de rezolvarea unei probleme minore de prezentare*. Situația pare absurdă, însă faptul că nu există (și nu poate exista) nici o soluție alternativă pentru problema de prezentare legitimează problema de interoperabilitate.

În sistemele de operare ale companiei Microsoft anterioare Windows Vista caracterele cu diacritice corecte sunt vizibile numai în Windows XP, și asta numai în anumite condiții.[14] Datorită cotei de piață uriașe a sistemelor de operare ale companiei Microsoft în rândul consumatorilor de conținut[15], *aceste considerente fac migrarea către caracterele cu diacritice corecte o chestiune discutabilă în absența penetrării masive a sistemelor de operare Windows Vista sau mai noi* pe piață.

Pe de altă parte, unii factori de decizie ai diverselor colecții de text de limbă română vor fi în mod inevitabil *early adopters*[16] ai noilor caractere cu diacritice corecte. Pe măsură ce trece timpul, pe măsură ce penetrează Windows Vista și sisteme de operare mai recente și pe măsură ce diverse colecții de text migrează la noile diacritice, va crește masa de conținut și de consumatori de conținut axați pe noile caractere. Odată ce se atinge o masă critică, *acei creatori de conținut care vor mai oferi text cu diacriticele incorecte vor fi văzuți ca depășiți*. Dacă în prezent există motive justificate pentru a amâna migrarea către diacriticele corecte[17], odată ce se atinge masa critică *nu va mai exista nicio scuză pentru întârzierea migrării*: în ultimă instanță, diacriticele noi sunt cele corecte, iar cele vechi sunt pur și simplu incorecte în limba română!

Totuși voi arăta mai jos că independent de felul în care sunt văzuți din afară sau de corectitudinea tehnică a diacriticelor folosită în colecțiile lor de text, unii dintre creatorii de conținut de limbă română vor avea *motive întemeiate pentru a adopta noile diacritice înaintea celorlalți*, iar alții vor avea *motive întemeiate pentru a întârzia migrarea pentru o perioadă semnificativă* chiar și după momentul apariției masei critice. Scopul acestui

[12] Există totuși situații în care diferența este ușor de sesizat chiar și pentru un consumator neavizat, mai ales atunci când se folosesc mărimi mari de literă (titlul unei cărți pe copertă, titlurile de pe afișe sau materiale publicitare etc.)
[13] Mai puțin cazul în care un text scris în română conține citate sau nume turcești. Chestiunea este discutată mai pe larg în secțiunea dedicată colecțiilor de text.
[14] Chestiunea este analizată pe larg în secțiunea dedicată consumatorilor de conținut.
[15] Vezi Tabelul 1
[16] Persoane (fizice sau juridice) care doresc să adopte cât mai repede tehnologiile cele mai recente.
[17] Pentru brevitate voi folosi în continuare sintagmele „diacritice corecte" și „diacritice noi" pentru „caractere care folosesc semnele diacritice corecte" (virgule), respectiv „diacritice incorecte" și „diacritice vechi" pentru „caractere care folosesc semnele diacritice vechi" (sedile).

document este tocmai acela de a identifica factorii care influențează momentul în care trebuie luate aceste decizii, de la caz la caz.

4. Factori de influență

După cum am arătat în secțiunea , două forțe opuse acționează simultan asupra deciziei de migrare la diacriticele corecte:

- *Pentru migrare cât mai rapidă*: tehnologia este deja disponibilă, conținutul ar putea fi deja vizualizat de majoritatea consumatorilor, iar rezultatul ar fi utilizarea caracterelor corecte în limba română. În plus, imaginea ultimilor creatori de conținut care să migreze va avea probabil de suferit într-o oarecare măsură.
- *Pentru amânarea migrării*: diverse probleme de lizibilitate și interoperabilitate, dintre care unele foarte semnificative.

Aceste două forțe vor avea o evoluție dinamică de-a lungul timpului, în sensul că prima va crește în defavoarea celei de-a doua, până la eliminarea completă a acesteia din urmă.

După o analiză îndelungată a structurii diverșilor factori care influențează aceste două forțe contrare, am identificat trei piloni pe care se sprijină întregul raționament:

a) *Consumatorii de conținut:* cititorii, utilizatorii produselor software etc.
b) *Creatorii de conținut*: edituri, deținători de site-uri, producători de software etc.
c) *Colecțiile de text*: conținutul efectiv al revistelor, site-urilor, aplicațiilor etc.

În acest capitol voi analiza felul în care fiecare dintre acești trei piloni afectează fiecare dintre cele două forțe identificate mai sus.

4.1 Consumatorii de conținut

Este de la sine înțeles că cel mai important dintre cei trei piloni este cel reprezentat de consumatorii de conținut: indiferent de capabilitățile tehnice ale creatorilor de conținut și de colecțiile lor de text, orice demers este inutil în măsura în care conținutul nu poate fi consumat sau, în cazul aplicațiilor interactive, consumatorul nu poate interacționa cu interfața în așa fel încât să obțină acces la conținutul propriu-zis.

Prima întrebare în ceea ce-l privește pe consumatorul de conținut este legată de modalitatea prin care acesta consumă în mod concret conținutul. *Dacă mediul final de consum nu implică tehnologii aflate sub controlul consumatorului, atunci consumatorul nu este un factor semnificativ* în luarea deciziei de migrare. În această situație se află conținutul prezentat exclusiv pe medii tipărite, sau în general pe orice medii în care consumatorul are un rol pasiv din punctul de vedere al tehnologiilor implicate (cărți, jurnale, reviste, afișe, prezentări video, filme și așa mai departe). În plus, conținutul consumat în condiții controlate de creatorii de conținut beneficiază în mare măsură de aceleași derogări (de exemplu aplicații care rulează în mod chioșc[18], aplicații online sau client-side care rulează într-un mediu proprietar, conținut prezentat pe hardware dedicat precum cititoarele de cărți electronice).

[18] Terminale dedicate, instalate în locuri publice, așa cum sunt cele din aeroporturi, gări, puncte de interes turistic etc.

Dacă însă mediul de consum este dependent de tehnologii controlate de consumator, atunci devin relevante două subcategorii de factori care influențează capacitatea consumatorului de a utiliza conținutul:[19]

1. *Lizibilitatea* – pot consuma conținutul?[20]
2. *Interactivitatea* – pot interacționa cu interfața?

Pentru a cântări capacitatea de lizibilitate a consumatorilor în contextul diacriticelor corecte trebuie analizat nivelul de utilizare al platformelor care suportă diacriticele corecte *în contextul consumatorilor colecției de text* în speță.[21]

Datele statistice utilizate în acest document sunt următoarele:[22]

Tabelul 1: Datele utilizate pentru generarea graficelor

Sistem de operare	**Cota de piață**	**Afișează**	**Scrie**[23]
Windows XP[24]	73,14%	Lizibil (70,8%)	Simplu (34%)
		Ilizibil (29,2%)	Dificil (66%)
Windows Vista	8,91%	Perfect	Simplu
Windows 7	15,89%	Perfect	Simplu
Mac OS X	0,75%	Perfect	Simplu
Linux	0,74%	Nesigur	Simplu
Altele	0,73%	Nesigur	Dificil

Tabelul 2: Cota de piață a diverselor versiuni de Internet Explorer

Versiune Internet Explorer	**Cota de piață**
Internet Explorer 8	17,95%
Internet Explorer 7	9,75%
Internet Explorer 6	11,43%
Altele	0,11%

[19] Cel mai reprezentativ exemplu în acest sens sunt colecțiile de text ale site-urilor și aplicațiilor Internet/intranet. În aceeași situație se află însă orice colecție de text care poate fi interpretată pe mai multe platforme software – aplicații, documente distribuite, mesaje e-mail (e.g. newsletters) și așa mai departe.

[20] Includ în această subcategorie și problemele de accesibilitate pentru persoanele cu dizabilități – de exemplu în ce măsură aplicațiile care citesc textul pentru consumatorii nevăzători sunt capabile să recunoască și să interpreteze corect texte scrise cu diacriticele noi.

[21] Această precizare este crucială pentru o analiză corectă în cazul unor situații particulare; vezi și nota .

[22] Datele din Tabelul 1 și Tabelul 2 sunt colectate în mai 2010 de la http://gs.statcounter.com/, pentru consumatorii din România. Pentru o analiză riguros exactă a consumatorilor de conținut în limba română la nivel global ar fi necesare datele statistice asociate acestui grup demografic specific, independent de poziția geografică (consumatorii de conținut în limba română nu se găsesc numai în România, ci și în Republica Moldova și pe alte meridiane; în plus, nu toți consumatorii de conținut din România preferă limba română). Totuși, așa după cum am indicat în mod repetat în acest document, fiecare creator de conținut trebuie să țină cont de capabilitățile tehnice ale propriilor consumatori. Prin urmare am considerat acceptabile aproximările rezultate din utilizarea unui criteriu geografic pentru ilustrarea orientativă a situației curente în acest document.

[23] Statistică relevantă în special pentru secțiunea legată de creatorii de conținut.

[24] Aproximări semnificative în ambele subcategorii. Fracțiile din subcategorii sunt procente din cota Windows XP.

În privința Windows XP, care deocamdată rămâne lider detașat (nu numai că este cea mai utilizată platformă, dar este mai utilizată decât toate celelalte la un loc), situația este la fel de incertă, însă merită investigată. La instalare, Windows XP nu suportă deloc diacriticele corecte – pur și simplu pe ecran apar niște „pătrățele" în locul caracterelor respective (ultima coloană din Tabelul 3). Există două modalități principale de a obține compatibilitate cu diacriticele corecte în Windows XP[25]:

- Prin instalarea explicită a unui pachet software suplimentar de la Microsoft[26].
- Prin instalarea Internet Explorer 7 sau Internet Explorer 8, aplicații care sunt în mod normal instalate prin mecanismul de actualizate automată al Windows.

Pentru prima opțiune nu avem la dispoziție statistici, însă este destul de puțin probabil că un consumator oarecare de conținut a făcut efortul să descarce și să instaleze un astfel de pachet software. Pe de altă parte, actualizarea navigatorului Internet Explorer este una automată (și dezirabilă pentru motive independente de diacritice), deci este de așteptat ca o parte semnificativă a consumatorilor să fi instalat această actualizare.

Figura 2: Gradul de utilizare al diverselor sisteme de operare

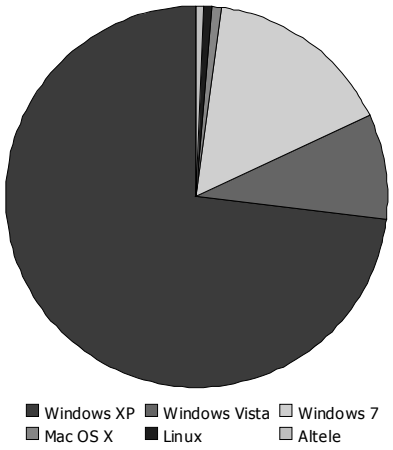
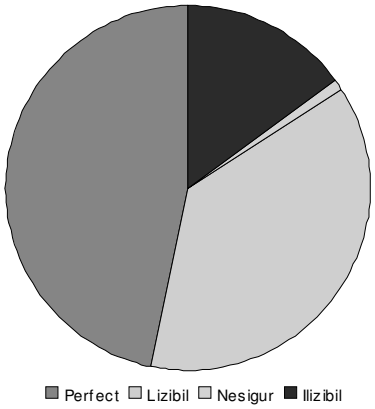


Figura 3: Capacitatea sistemelor de operare de a afișa diacriticele



[25] http://www.microsoft.com/Romania/Diacritice.aspx – în realitate atât Internet Explorer 8 cât și Internet Explorer 7 sunt capabile să opereze substituția de corp de literă pentru a obține rezultatele de pe coloana a doua din Tabelul 3.
[26] „Actualizare de fonturi corespunzătoare extinderii Uniunii Europene" de la http://www.microsoft.com/downloads/details.aspx?FamilyID=0ec6f335-c3de-44c5-a13d-a1e7cea5ddea&DisplayLang=ro

Totuşi asta înseamnă că mai bine de o treime dintre utilizatorii de Internet Explorer nu pot vedea text scris cu diacriticele corecte, atâta timp cât utilizează Windows XP. Nu avem la dispoziţie date statistice globale de încredere care să coroboreze sistemul de operare cu navigatorul utilizat[27], deci vom fi nevoiţi să presupunem că toţi utilizatorii de Windows XP se supun aceleiaşi proporţii identificate pentru Internet Explorer, indiferent dacă utilizează acest navigator sau altul.

Deja în 2010 *majoritatea consumatorilor de conţinut pot să citească text care foloseşte diacriticele corecte*. Există totuşi câteva rezerve semnificative:

- Utilizatorii pentru care textul este doar lizibil (nu perfect) pot citi textul, însă, în funcţie de situaţie, *unii vor observa o diferenţă sesizabilă, inestetică şi deranjantă de afişare a caracterelor respective* (a doua coloană din Tabelul 3). Suportul pentru diacriticele corecte se limitează în Windows XP la numai câteva corpuri de literă – pentru celelalte, sistemul de operare substituie pur şi simplu caracterele respective cu aceleaşi caractere din cele mai apropiate corpuri de literă pentru care dispune de caracterele în speţă. Rezultatul este cel aşteptat: textul este lizibil, însă în funcţie de diferenţa vizuală dintre corpul de literă utilizat în text şi cel disponibil rezultatele pot fi inestetice (în tabel am folosit Arial Black).
- Chiar în cazul colecţiilor de text accesibile via Internet este posibil ca unele aplicaţii specifice să se adreseze în mod particular consumatorilor care utilizează o paletă relativ îngustă de sisteme de operare. Un exemplu relevant sunt aplicaţiile sau subdomeniile dedicate pentru platforme mobile.[28]

	Windows 7, Vista	Windows XP (nou)	Windows XP (vechi)
Diacritice corecte	arşiţă	arşiţă	ar□i□ă
Diacritice vechi	arşiţă	arşiţă	arşiţă

Tabelul 3: Afişarea celor două tipuri de diacritice în funcţie de gradul de lizibilitate al diacriticelor noi (de notat că diacriticele vechi sunt perfect lizibile indiferent de context)

[27] De fapt statistica ideală ar fi chiar cea pe care încerc să o estimez aici (lizibilitate perfectă/lizibilitate limitată/ilizibilitate); în lipsa ei, cea mai bună aproximare ar fi o coroborare a lizibilităţii limitate (gradul de instalare al EUupdate.EXE sau IE7 sau IE8) cu sistemul de operare Windows XP. Totuşi procentele sunt deocamdată suficient de mari în toate categoriile încât aproximările din text să nu afecteze în mod semnificativ calitatea analizei.

[28] Multe site-uri publice oferă alternative pentru platforme mobile. De pildă un consumator care accesează pentru prima dată pagini din domeniul http://www.moongate.ro/ folosind un dispozitiv mobil este redirecţionat automat către pagina corespunzătoare din domeniul http://m.moongate.ro/; în cazul acestui al doilea domeniu sunt relevante numai statisticile legate de platformele mobile.

Acestea fiind spuse, trebuie menționat în mod proeminent că indiferent de platforma specifică a consumatorilor, indiferent de publicul țintă al colecției de text și indiferent de numărul lor, *fracțiunea consumatorilor care nu pot vizualiza diacriticele corecte se vor afla practic în imposibilitate de a consuma conținutul.* Iar alternativa este afișarea aceluiași text, utilizând caractere cu semne diacritice tehnic incorecte, dar care sunt aproape identice din punct de vedere vizual și pe care le poate citi oricine (vezi al doilea rând din Tabelul 3). În ultimă instanță trebuie pusă în balanță corectitudinea academică față de pierderea de facto a unei fracțiuni a cititorilor.

Celălalt factor care trebuie luat în considerare este capacitatea consumatorilor de a interacționa cu interfața colecției de text. Cea mai proeminentă funcție a interfeței în această privință este funcționalitatea de căutare: dacă colecția de text conține diacritice corecte iar consumatorul operează o căutare utilizând diacriticele vechi (sau viceversa)[29] atunci consumatorul nu va obține rezultatele dorite. *Practic toate problemele de interacțiune pot fi rezolvate prin soluții tehnice*, însă acestea trebuie luate în considerare și rezolvate din timp.[30] Totuși *problemele de interacțiune sunt printre puținele care pot fi rezolvate înainte de începerea migrării și* ***ar trebui rezolvate cât mai curând***.[31]

4.2 Creatorii de conținut

Al doilea pilon de influență al deciziei de migrare este capacitatea creatorilor de conținut de a crea conținut folosind diacriticele corecte. *Este prea puțin important dacă cititorii pot citi, atâta vreme cât scriitorii nu pot scrie.*

Prin urmare factorii care influențează creatorii de conținut sunt dictați în primul rând de o logică similară celei legate de consumatori:

> Poate autoritatea sub egida căreia se generează conținut să controleze mijloacele tehnice ale creatorilor individuali de conținut?

Răspunsurile posibile la această întrebare sunt cu mult mai variate decât în cazul consumatorilor. Am ales aici numai cazurile extreme, pentru exemplificare:

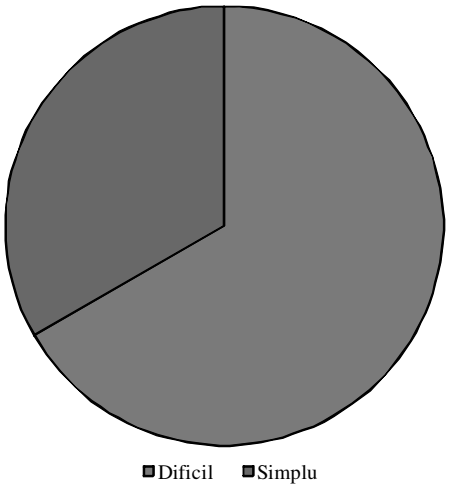


Figura 4: În ce masura se poate scrie text cu diacriticele corecte

[29] Vezi și ultima coloană din Tabelul 1, reprezentată grafic în Figura 3.

[30] Aproape orice problemă de acest fel poate fi rezolvată prin interpretarea flexibilă a datelor de intrare, așa cum procedează Google sau DEX online.

[31] O colecție de text proeminentă de limbă română care a migrat foarte curând la diacriticele corecte este DEX online (http://dexonline.ro/). Au fost însă luate în calcul aproape toate problemele identificate în acest document: consumatorii care nu pot citi diacriticele corecte pot alege să consulte dicționarul folosind diacriticele vechi, iar interacțiunile cu colecția de text ignoră în mod implicit semnele diacritice din textul de intrare. Singura scăpare este legată de limba turcă, în sensul că și cuvintele care ar trebui scrise cu sedilă au fost convertite automat la forma cu virgulă (e.g. http://dexonline.ro/definitie/siret versus http://tr.wiktionary.org/wiki/%C5%9Ferit). Totuși absența unei necesități de interoperabilitate ulterioară internă dintre dicționar și orice altă colecție de text face ca această scăpare să fie lipsită de orice efecte adverse concrete.

- *Redactori tradiţionali* (jurnal, revistă, carte tipărită ş.a.m.d.): aceşti redactori lucrează la sediul societăţii care i-a angajat, iar societatea respectivă are control complet asupra platformei software utilizate la sediu. Chiar şi redactorii care preferă să folosească mijloace tehnice proprii pot fi constrânşi să urmeze standardele societăţii (e.g. atunci când redactează text utilizând laptopul sau computerul propriu). În acest caz creatorii de conţinut sunt un factor neglijabil, deoarece autoritatea decizională va opera migrarea pe baza celorlalţi doi piloni.
- *Redactori voluntari, independenţi* (de exemplu redactori la Wikipedia sau alte proiecte similare, persoanele care contribuie cu comentarii în diverse site-uri, site-uri de socializare ş.a.m.d.): acesta este unul dintre cei mai puternici factori pentru amânarea adoptării diacriticelor corecte (vezi dedesubt).

După cum se vede în Figura 4, *mai bine de jumătate dintre creatorii independenţi de conţinut au dificultăţi în a scrie text utilizând diacriticele corecte*.[32] Motivul central pentru această limitare este faptul că utilizatorii de Windows XP, deocamdată majoritari în statisticile globale, nu pot genera conţinut utilizând diacriticele corecte decât în condiţii destul de stricte.[33] În aceste condiţii decizia trebuie luată în concordanţă cu capacitatea factorilor decizionali de a influenţa mijloacele tehnice utilizate de creatorii individuali de conţinut, sau de a implementa soluţii tehnice pentru ameliorarea sau rezolvarea acestei probleme specifice. În cazul proiectelor bazate pe voluntariat, aşa cum este Wikipedia, este evident că migrarea în condiţiile actuale nu ar avea sorţi de izbândă, chiar lăsând la o parte ceilalţi piloni, în absenţa unor soluţii tehnice specifice.[34]

4.3 Colecţia de text

Însăşi colecţia de text, cel mai puţin important dintre cei trei piloni identificaţi aici, va fi pentru mulţi creatori de conţinut impedimentul major în decizia de migrare, chiar şi atunci când migrarea ar fi dezirabilă din celelalte puncte de vedere. Orice analiză a deciziei de migrare către diacriticele corecte este în mod firesc axată în primul rând pe utilizabilitate din punctul de vedere al consumatorului de conţinut, în al doilea rând pe capacitatea creatorului de conţinut de a genera conţinut şi abia în ultimul rând pe *disponibilitatea creatorului de conţinut de a migra în mod retroactiv conţinutul existent la diacriticele corecte*.

Colecţia de text influenţează decizia de migrare în funcţie de următorii factori:

1. *Relevanţă*: este relevant să vă puneţi întrebări legate de migrarea colecţiei de text numai în măsura în care aceasta este vizibilă. De exemplu arhiva digitală a unui periodic publicat exclusiv în formă tipărită nu face în general obiectul migrării. Evident, dacă există o parte accesibilă în format electronic şi una stocată intern atunci numai partea accesibilă este relevantă.
2. *Mărime*: semnificaţia acestui factor trebuie coroborată în general cu următorul, complexitatea colecţiei de text. Există însă un caz particular în care contează exclusiv mărimea: atunci când ea este nulă. Dacă un creator de conţinut începe

[32] Datele sunt extrase de pe ultima coloană din Tabelul 1.
[33] http://www.stefamedia.ro/diacritice-romanesti-corecte-in-windows-xp/
[34] La Wikipedia s-a luat deja decizia migrării la diacriticele corecte, deoarece au fost identificate soluţii tehnice concrete care permit vastei majorităţi a redactorilor individuali să genereze conţinut utilizând diacriticele corecte indiferent de platforma software pe care o utilizează. Pentru detalii despre acest caz particular vezi secţiunea dedicată studiului de caz Wikipedia.

lucrul la o colecție de text complet nouă atunci trebuie cântărită cu atenție opțiunea de a utiliza de la bun început diacriticele corecte – în măsura în care aceasta este o opțiune acceptabilă din celelalte puncte de vedere atunci ar trebui adoptată, întrucât astfel se va evita mai târziu costisitorul proces de migrare retroactivă.

3. *Complexitate*: am văzut mai sus că distincția dintre caracterele cu virgulă și cele cu sedilă nu are valoare semantică în limba română. Altfel spus, un text scris în exclusivitate în limba română ar putea fi convertit cu ușurință la diacriticele corecte printr-o simplă operațiune de căutare și înlocuire automată. În general acest lucru este adevărat, însă cu unele rezerve pe care le voi analiza dedesubt.

Complexitatea colecției de text este o măsură a frecvenței situațiilor care necesită intervenție umană. În cazul colecției de text de la Wikipedia am identificat câteva tipologii specifice de situații problematice, probabil reprezentative pentru cazul general:

- *Texte scrise în altă limbă* (cel mai notabil în turcă) fără notație adecvată.[35] Dacă textele (inclusiv numele de persoane, locuri, evenimente ș.a.m.d.) în turcă sunt marcate explicit ca fiind scrise în turcă există posibilitatea de a automatiza procesul prin evitarea schimbării semnelor diacritice în cazul acestor fragmente de text.
- *Identificatori de resurse care conțin semne diacritice*. Pe lângă cazul general (URI) mai pot exista o sumedenie de identificatori interni sau externi a căror integritate structurală trebuie menținută de-a lungul procesului de migrare, în ambele sensuri.[36]
- *Interoperabilitatea cu alte colecții de text*, în special în condițiile unei migrări parțiale.[37]

5. Studiu de caz: Wikipedia în limba română

Am fost implicat în aproape toate discuțiile de la Wikipedia în limba română în privința deciziei de migrare la diacriticele corecte, iar mărimea și complexitatea colecției de text, numărul mare și eterogenitatea redactorilor și consumatorilor de conținut ai acestui proiect îl fac probabil unul dintre cele mai bune studii de caz posibile în contextul acestui document.

La Wikipedia în limba română, discuțiile despre migrarea la diacriticele corecte au început încă din anul 2007. Deși la acel moment au fost considerate premature, în urma discuției a fost creată o pagină dedicată subiectului.[38] La sfârșitul aceluiași an a fost

[35] De exemplu în HTML: http://www.w3.org/TR/WCAG10-HTML-TECHS/#language [en]

[36] Cele mai la îndemână exemple sunt legate de colecția de text de la Wikipedia în limba română. Printre identificatorii interni de resurse se numără legăturile interne între articole și cele care leagă articole despre același subiect în mai multe limbi. Identificatorii externi conținuți în Wikipedia sunt legăturile (URI) către pagini de pe alte site-uri și care pot conține caractere cu diacritice. Identificatorii externi pe care trebuie să-i gestioneze Wikipedia sunt legăturile (URI) dinspre alte site-uri către articolele din Wikipedia care pot conține caractere cu diacritice; aceasta este mai degrabă o resposabilitate morală datorată numărului relativ mare de documente care fac trimitere la enciclopedie.

[37] De exemplu atunci când colecția de text are relevanță parțială în privința diacriticelor, dar partea publică a colecției de text (cea care urmează să fie migrată) interacționează prin sisteme automate cu partea istorică/privată/confidențială care nu este migrată.

[38] http://ro.wikipedia.org/wiki/Wikipedia:Corectarea_diacriticelor

implementată o soluţie tehnică de natură să forţeze utilizarea diacriticelor *vechi*, în scopul uniformizării conţinutului (unii redactori deja scriau conţinut folosind diacriticele noi). La sfârşitul lui 2008 şi începutul lui 2009 discuţiile au început să se orienteze către identificarea problemelor şi soluţiilor tehnice concrete asociate utilizării diacriticelor corecte şi s-au început câteva demersuri tehnice concrete, deşi fără efecte pentru consumatori, în vederea viitoarei migrări.

Dată fiind absenţa unor statistici şi analize concrete, comunitatea a decis la 1 martie 2010 să implementeze un proiect pilot prin care se permitea utilizarea diacriticelor corecte pe cele mai multe dintre paginile interne ale proiectului, cu scopul de a identifica numărul de redactori incapabili să le citească; nu s-a înregistrat nicio plângere.

Pe fondul discuţiilor de la Wikipedia şi în urma invitaţiei organizatorilor ConsILR de a participa la conferinţă, am organizat în perioada 27 aprilie–4 mai 2010 un sondaj naţional[39] în scopul identificării capabilităţilor tehnice ale consumatorilor de limbă română în privinţa lizibilităţii diacriticelor noi. Discuţiile din timpul sondajului au dus la identificarea unui criteriu concret pe baza căruia urma să se ia decizia migrării la diacriticele corecte.[40]

Sondajul a beneficiat de un nivel adecvat de expunere şi participare[41], dar am determinat ulterior că voturile exprimate au fost eronate în proporţie mult prea mare pentru orice scop practic.[42] Totuşi datele statistice obţinute prin analizarea vizitatorilor (ignorând voturile exprimate) au confirmat faptul că distribuţia capabilităţilor tehnice ale vizitatorilor Wikipedia în limbă română este foarte apropiată de datele statistice naţionale prezentate în acest document. Coroborând această informaţie cu soluţiile tehnice identificate anterior, de natură să amelioreze problema lizibilităţii pentru segmentul consumatorilor afectaţi, am constatat că a fost satisfăcut criteriul convenit în timpul desfăşurării sondajului. Prin urmare, pe 5 mai 2010 s-a luat decizia de migrare cât mai rapidă la diacriticele corecte.[43]

În cazul colecţiei de text de la Wikipedia în limba română procesul va fi cu siguranţă unul de durată, însă analiza şi execuţia până în acest moment au urmat o traiectorie foarte apropiată de cea ideală. La Wikipedia, calitatea analizei şi a deciziilor în această privinţă se datorează în cea mai mare măsură faptului că deciziile se iau prin consens, iar viziunile divergente ale diverşilor participanţi la discuţie au asigurat în cadrul discuţiilor o reprezentare optimă a celor două forţe identificate în secţiunea

6. Concluzii

Toate colecţiile de text care conţin text în limba română vor urma inevitabil, mai devreme sau mai târziu, standardul corect în privinţa semnelor diacritice. Singurul punct

[39] http://www.moongate.ro/products/diacritice/sondaj/

[40] „M*ai puţin de 3% de vizitatori pe care nu îi putem ajuta (nu văd diacritice sau văd majuscule şi au alt SO decât Windows sau Windows mai vechi de XP)*", la pagina menţionată în nota .

[41] http://www.moongate.ro/products/diacritice/sondaj/date.php#statistici

[42] Aproximativ 20% dintre respondenţii care au votat că nu pot citi diacriticele noi utilizau Windows Vista sau Windows 7, sisteme de operare despre care ştim cu siguranţă că în realitate le afişează perfect.

[43] http://ro.wikipedia.org/w/index.php?title=Discu%C5%A3ie_Wikipedia:Sfatul_B%C4%83tr%C3%A2nilor&oldid=3918861#Concluzii_finale

delicat este alegerea momentului optim pentru migrare. Diferența dintre cele două variante este mică din punct de vedere vizual, dar există dificultăți tehnice potențiale semnificative asociate migrării.

Am căutat să identific aici factorii care influențează alegerea pragmatică a momentului optim de migrare pe baza următoarei structuri:

- Consumatorii de conținut
  - în ce măsură tehnologia se află sub controlul consumatorului
  - în ce măsură pot consuma conținutul
  - în ce măsură pot interacționa cu interfața
- Creatorii de conținut
  - în ce măsură pot controla tehnologia utilizată de redactorii individuali
  - în ce măsură redactorii individuali pot crea text folosind diacriticele corecte
- Caracteristicile colecției de text
  - relevanța
  - mărimea
  - complexitatea

Creatorii de conținut de limbă română ar trebui să ia cât mai curând următoarele măsuri concrete:

- Implementarea măsurilor tehnice necesare pentru a permite consumatorilor care folosesc deja diacriticele corecte să interacționeze cu interfața (e.g. pentru funcțiile de căutare).
- Implementarea măsurilor tehnice necesare în vederea migrării, dacă este cazul (în particular identificarea explicită a textelor scrise în limba turcă).
- Analiza situației propriului caz particular, prin prisma acestui document.
- Identificarea factorilor critici aplicabili colecției de text analizate.
- Stabilirea unui criteriu concret pentru migrare, pe baza factorilor aplicabili.
- Planificarea migrării (metodologie tehnică, estimare buget și resurse necesare).
- Monitorizarea și previzionarea evoluției situației în raport cu criteriul ales, în așa fel încât să poată aloca resursele necesare concretizării migrării în timp util.

Ultima versiune a acestui document, precum și alte noutăți și resurse suplimentare în materie se găsesc la adresa http://www.moongate.ro/products/diacritice/

Mulțumiri *organizatorilor și referenților ConsILR*, pentru că mi-au oferit imboldul inițial și încurajarea ulterioară pentru cristalizarea acestui document în formă scrisă; domnului *Cristian Secară*, pentru informațiile concrete care m-au ajutat să corectez unele statistici care altfel ar fi fost cu siguranță greșite; domnului *Cristian Adam*, pentru notele legate de Latin-10; *colectivității Wikipedia* în limba română, pentru discuțiile întotdeauna deschise; în acest context le mulțumesc în particular redactorilor *Cezarika1* (primul care a ridicat această problemă și ne-a arătat de ce merită discutată), *Strainu* (un early adopter prin definiție, foarte capabil pe partea tehnică, cel care a împins constant în direcția adoptării cât mai rapide a diacriticelor corecte la Wikipedia și care m-a ajutat și cu sugestii în privința acestui document), *Danutz* (care mi-a indicat site-ul

http://gs.statcounter.com/, utilizat pentru toate statisticile din acest document; în versiuni mai vechi ale acestui document am folosit statistici globale, semnificativ diferite), *AdiJapan* (ca întotdeauna, a reușit să găsească un echilibru între interlocutorii mai tehnici și cei mai puțin tehnici, între cei avangardiști și cei conservatori); și, independent de Wikipedia, *tatălui meu și prietenilor* care m-au ajutat să dau forma curentă acestei lucrări (știți voi cine sunteți).

CONSTRUCŢIA AUTOMATĂ DE CORPUSURI MULTILINGVE

TIBERIU BOROŞ, DAN TUFIŞ, ALEXANDRU CEAUŞU

Institutul pentru Cercetarea Inteligenţei Artificiale – Academia Română

{tibi,tufis,aceausu}@racai.ro

Rezumat

Insuficienţa resurselor lingvistice pentru multe dintre limbile naturale este principalul impediment în progresul tehnologic al prelucrării automate a acestor limbi. Succesul abordărilor statistice în traducerea automată pentru limbile de largă circulaţie, bine echipate sub raport cantitativ şi calitativ cu resurse lingvistice, a evidenţiat încă o dată necesitatea colectării şi preprocesării corespunzătoare a unor volume cât mai mari posibile de resurse lingvistice. Pentru aplicaţiile multi- şi cros-linguale corpusurile paralele şi (mai recent) cele comparabile sunt resurse primare indispensabile. Articolul de faţă prezintă o serie de unelte ce pot fi folosite în extragerea automată de corpusuri paralele sau puternic comparabile.

1. Introducere

În ultimii 15-20 de ani abordările bazate pe tehnici inductive şi volume mari de date, ca şi creşterea spectaculoasă a performanţelor de calcul şi de stocare ale noilor generaţii de calculatoare au condus la progrese greu de anticipat în anii de început ai lingvisticii computaţionale şi ai traducerii automate. Utilizarea web-ului ca resursă primară de date, fără nici un fel de prelucrare lingvistică majoră, a demonstrat că utilizarea tehnologiilor statistice aplicate unor volume foarte mari de texte poate fi un răspuns la marea provocare a traducerii automate între toate limbile documentelor din spaţiul web. Google translate[1] deja oferă traducere automată pentru 52 de limbi (2652 de perechi limbă sursă-limbă ţintă). Sistemul Bing Translator[2] de la Microsoft oferă traducere automată pentru 30 de limbi (870 de perechi limbă sursă-limbă ţintă). Pe lângă aceste exemple de sisteme publice foarte cunoscute pot fi amintite şi companii mai noi[3], focalizate pe traducere automată care dezvoltă sisteme comerciale, asigurând însă pentru unele perechi de limbi o calitate a traducerii aproape de nivelul traducerii umane[4].

Dacă metodele de prelucrare numerică bazate exclusiv pe texte neprelucrate lingvistic (în engleză acest tip de prelucrare este numit *number crunching NC*) sunt aplicabile pentru orice pereche de limbi, în schimb volumul datelor necesare pentru o calitate acceptabilă a traducerii este uriaş (pentru unele perechi de limbi nici măcar conţinutul actual al web-ului nu este suficient), necesitând resurse de calcul şi stocare ce depăşesc dotările grupurilor obişnuite de cercetare. Numai marile companii îşi pot permite dezvoltări de sisteme de prelucrare multilingve pentru un număr mare de perechi

[1] http://translate.google.com/# (consultare la data de 10 aprilie 2010)
[2] http://www.microsofttranslator.com/Default.aspx (consultare la data de 10 aprilie 2010)
[3] http://www.translationguide.com/translation_company_links.php (o listă parţială la data de 10 aprilie 2010)
[4] http://www.languageweaver.com/page/home/

arbitrare de limbi. Pe de altă parte cercetări recente (Koehn et al., 2007), (Hoang et al.,2009), (Tufiș et al., 2009) ș.a. au arătat că atunci când textele sunt prelucrate lingvistic (segmentate, dezambiguizate morpho-lexical, lematizate, parsate) cu un volum de date mult mai mic decât conținutul (pentru o anumită pereche de limbi) întregului web se pot obține traduceri comparabile și chiar superioare metodelor de tip NC. Fără îndoială însă că și pentru astfel de abordări cu cât volumul datelor primare este mai mare calitatea traducerilor crește.

Pentru sistemele de traducere automată bazate pe metode statistice (indiferent dacă sunt de tip NC sau includ prelucrări lingvistice) cele mai valoroase resurse textuale sunt corpusurile paralele ce conțin texte în două sau mai multe limbi astfel încât ele reprezintă traduceri reciproce. Unul dintre primele corpusuri paralele bilingv (engleză-franceză) este corpusul Hansards conținând transcrieri ale dezbaterilor în Parlamentul Canadei în perioada 1975-1988 (Germann, 2001). El a constituit întotdeauna o resursă fundamentală pentru studiile cros-linguale pentru perechea de limbi engleză-franceză.

Multilingvismul este o caracteristică esențială a Europei și implicațiile culturale, sociale și economice au ridicat multilingvitatea la nivelul unei preocupări politice de prim rang și absolut toate instituțiile Uniunii Europene produc un volum foarte mare de documente paralele. Necesitatea accesului cercetătorilor din domeniul prelucrării limbajelor naturale la aceste documente a determinat factorii responsabili (de exemplu OPOCE – biroul pentru publicații oficiale ale Uniunii Europene) să facă publice formatul electronic al unui volum din ce în ce mai mare de documente paralele pe baza cărora cercetătorii au construit primele corpusuri paralele europene publice. Primul dintre acestea a fost EuroParl (Koehn, 2005), disponibil pentru 11 din limbile comunității europene încă din 2001. El conține transcrieri ale sesiunilor din Parlamentul European din perioada 1996-2001. Versiunile următoare[5] au extins cantitatea de texte, dar nu și numărul de limbi. Câțiva ani mai târziu, în 2006, a fost creat și distribuit prima variantă a corpusului JRC-Acquis (Steinberger et al., 2006). Versiunea V3 este în prezent[6] cel mai mare corpus multilingv disponibil, conținând texte de natură juridică în 22 de limbi ale Uniunii Europene. Acest corpus conține peste 1 miliard de cuvinte, în medie 48 milioane de cuvinte/limbă. Din perspectiva limbii române, alături de JRC-Acquis mai sunt relevante corpusurile EMEA și OpenSubs (Tiedemann, 2009). EMEA conține documente ale Agenție Europene a Medicamentului în 22 de limbi, iar OpenSubs conține subtitrări în 30 de limbi. Textele românești din JRC-Acquis și EMEA conțin diacritice în schimb ele lipsesc din OpenSubs ceea ce face acest ultim corpus mai puțin util pentru prelucrarea limbii române.

La o privire generală, cele mai multe corpusuri paralele au două mari probleme inerente: conțin, în principal, limbi de largă circulație și sunt specifice unui anumit domeniu.

Folosirea de corpusuri comparabile devine o alternativă viabilă pentru antrenarea sistemelor de traducere în cazul domeniilor și limbilor care au o mică reprezentare în corpusuri paralele. Un corpus comparabil reprezintă o colecție de texte în două sau mai

[5] Versiunea v5, distribuită la începutul anului 2010, acoperă perioada 1996-2009 și conține aproximativ 55 milioane de cuvinte pentru fiecare din cele 11 limbi: franceză, italiană, spaniolă, portugheză, engleză, olandeză, germană, daneză, suedeză, greacă și finlandeză.
[6] http://langtech.jrc.it/JRC-Acquis.html

multe limbi care deşi nu reprezintă traduceri reciproce riguroase, conţin informaţii similare. Aceste tipuri de corpusuri au grade de comparabilitate diferite (Skandina et al., 2010), cele mai utile fiind cele clasificate drept „puternic comparabile" (eng. *strongly comparable*). Aceste corpusuri comparabile conţin informaţii despre aceleaşi lucruri, folosesc acelaşi registru lingvistic şi au un grad ridicat de suprapunere la nivelul echivalenţilor lexicali de traducere. Deşi antrenarea sistemelor de traducere folosind corpusuri comparabile implică un nivel de complexitate ridicat faţă de antrenarea tradiţională ce foloseşte resurse paralele, textele comparabile sunt disponibile în proporţie mult mai mare decât cele paralele.

2. Combinarea modelelor de traducere extrase din corpusuri paralele şi corpusuri comparabile

În cadrul proiectului naţional STAR (PNII – IDEI 742/19.01.2009) la Institutul de Cercetări pentru Inteligenţa Artificială a fost realizat prototipul unui sistem de traducere pentru perechea de limbi engleză-română (Ceauşu, 2009) folosind platforma Moses (Koehn et al, 2007) şi serviciile de procesare a textului (Tufiş et al, 2007). Sistemul a fost antrenat pe un corpus de aproape un milion de unităţi de traducere cu peste treizeci de milioane de atomi lexicali. Corpusul este compus în proporţie de 75% din texte din domeniul juridic, 5% texte din domeniul jurnalistic. Restul de 20% sunt echivalenţi de traducere şi unităţi de traducere extrase din ontologia lexicală românească Ro-Wordnet extinsă cu terminologie juridică (Tufiş et al, 2008). Experimentele noastre au arătat că traducerea textelor din domeniul juridic are o calitate bună dar, deşi inteligibile, traducerile textelor din alte domenii (de ex. sport, medicină, turism etc.) au o calitate mult mai slabă. O soluţie raţională pentru remedierea acestui aspect constă în construcţia mai multor modele statistice de limbă şi de traducere, fiecare caracteristic unui anumit registru textual. În continuare, un modul preliminar ar putea clasifica un text nou ce urmează a fi tradus într-una din clasele pentru care există modele de traducere şi sistemul va efectua traducerea folosind modelul specific. O altă variantă constă în a combina diferitele modele, atribuind dinamic, în funcţie de domeniul textului ce urmează a fi tradus, ponderi de influenţă diferite modelelor statistice combinate.

Proiectul european Accurat[7] (248347/FP7), lansat la începutul acestui an, are ca obiectiv, printre altele, construcţia unor corpusuri comparabile cât mai mari pentru care să poată complementa puţinele corpusuri paralele disponibile pentru limbile proiectului (inclusiv româna). Combinarea modelelor de traducere construite din cele două categorii de corpusuri se va face studiind variantele amintite anterior şi alegând soluţia cea mai performantă.

Pentru a extinde sistemul de traducere şi pentru alt domeniu decât cel juridic, decodorului Moses îi pot fi adăugate modele de traducere antrenate folosind corpusurile comparabile.

Moses este un decodor pentru sistemele de traducere automată care extinde traducerea formei de ocurenţă a cuvintelor cu modele de traducere factorizate. Spre deosebire de decodarea bazată pe echivalenţii de traducere constituiţi din secvenţe contigue de cuvinte, care se bazează numai pe forma de ocurenţă a cuvintelor, traducerea factorizată

[7] http://www.accurat-project.eu/

poate lua în considerare informații suplimentare asociate atomului lexical, cum ar fi partea de vorbire, forma de dicționar a cuvântului sau descrierea sa morfo-sintactică.

În conformitate cu modelul traducerii factorizate, procesul de traducere presupune căutarea traducerii *t* (în limba țintă) care maximizează o combinație liniară a probabilităților diverșilor factori utilizați. Probabilitatea de traducere și regula de decizie este dată de:

$$t^* = \arg\max_{e} \sum_{k=1}^{n} \lambda_k h_k(t, s) \quad (1)$$

unde h_k (t,s) este unul din cei *n* factori (o funcție de trăsături caracteristice perechii <t,s>) iar λ_k este ponderea acestuia. Cei mai importanți factori care contribuie la probabilitatea de traducere provin din modelele de traducere, modelele de limbă, modelele de generare și cele de distorsiune. Combinarea modelelor de traducere din corpusuri comparabile și a modelelor de traducere din corpusuri paralele este dată de ponderile λ_k folosite pentru fiecare factor.

Stabilirea ponderilor λ_k se realizează prin proceduri de învățare automată. Una dintre cele mai utilizate proceduri este MERT (Minimal Error Rate Training) (Och, 2003) care presupune existența unor traduceri de referință, în raport cu care se induc valorile λ_k pentru care diferența dintre textul de tradus și textul de referință este minimă în raport cu o măsură de similaritate. Algoritmul MERT este un proces iterativ, computațional foarte intens, dar complet nesupervizat. Decodorul MOSES este însoțit în distribuția standard și de programul care implementează algoritmul MERT care implicit utilizează scorul BLEU (Papineni, et al., 2002). Această implementare (Bertoldi et al, 2009) a algoritmului MERT poate folosi opțional și alte măsuri de evaluare.

Pentru colectarea corpusurilor necesare proiectelor noastre a fost construit un program specializat care va fi descris în continuare. După cum se va vedea din prezentarea sistemului, acesta poate colecta atât corpusuri paralele cât și corpusuri comparabile, în funcție de natura conținutului multilingv al site-urilor prelucrate.

3. Descrierea sistemului de colectare a corpusurilor multilingve

O componentă importantă a motoarelor de căutare este un modul (numit *spider* sau *crawler*) care citește fiecare pagină și reține legăturile care pleacă din aceasta (Kobayashi and Takeda, 2000). Astfel se generează un graf plecând de la o listă inițială de adrese ce conțin referințe către alte locații. Procesul de generare poate fi automat (prin liste de căutare generate la prima indexare urmând ca la anumite intervale de timp să se verifice schimbări în conținutul acestora), manual (lista de site-uri este actualizată prin intervenție umană) sau hibrid.

Problema care apare în cazul extragerii de corpus este relevanța rezultatelor. În multe situații se obțin și referințe către site-uri fără nici o legătură cu tema de la care s-a plecat și pentru care nu se mai pot aplica aceleași reguli de indexare.

De exemplu, pe site-ul Wikitravel la articolul Bucureşti se regăsesc următoarele referinţe ce nu prezintă interes pentru extragerea de corpus, deoarece paginile respective nu se încadrează în tiparul acceptat de aplicaţie:

- **Carsrent** – există doar versiunea în limba engleză
- **Allrent, Budget** s – linkurile-ul sunt invalide

Din acest motiv, listele trebuie supuse unui proces de filtrare în funcţie de specificul fiecărei pagini. Stabilirea link-urilor ce vor fi verificate se face prin filtrarea acestora conform unor criterii bine alese. De exemplu:

– Dacă se doreşte ca articolul/documentul vizitat să facă parte din pagina curentă se poate recurge la eliminarea elementelor ***href*** ce conţin ***http://*** în interior (majoritatea referinţelor interne se fac prin calea relativă către document) sau eventuala eliminare a legăturilor ce nu conţin adresa de bază a site-ului curent.
– Există cazuri în care se poate face o selectare bazată pe organizarea internă a paginii web. De exemplu, toate paginile de ştiri conţin în adresele lor ***/news/*** iar toate paginile referitoare la vreme conţin în adresele lor ***/weather/***

Indiferent de metoda care se alege, dacă structura sitului se modifică la un moment dat, aplicaţia trebuie actualizată corespunzător, atrăgând după sine modificări în codul sursă, care devine greu de citit şi reutilizat în alte aplicaţii.

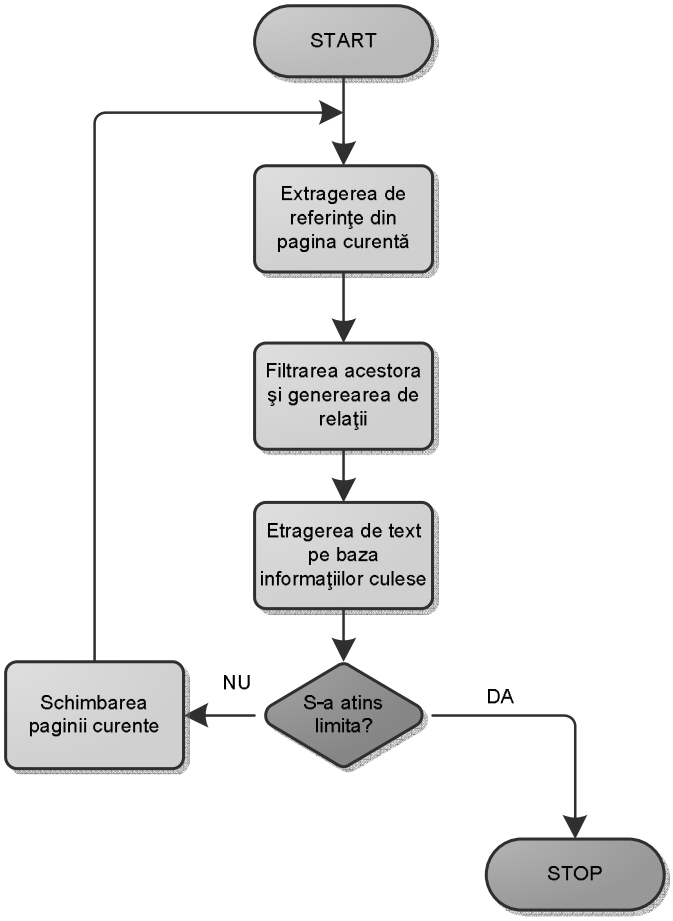


Figura 1: Diagrama de flux pentru algoritmul de extragere de text din pagini web

În continuare propunem o soluție de simplificare și minimizare a efortului depus la adăugarea unor noi criterii de indexare sau modificarea celor deja existente, prin utilizarea unei arhitecturi bazate pe diagrame de flux. În (Chappell, 2009) sunt prezentate avantajele unei soluții de acest tip. Astfel, fiecare bloc de procesare al diagramei de flux este privit independent de celelalte putând fi editat în orice moment și permițându-se adăugarea de pași suplimentari la fluxul deja existent. În fundal (eng. *background)* sunt lansate procese de tip consolă care primesc ca parametri de intrare fluxul și generează ieșire de tip text. Aceasta este preluată și este transmisă, în urma aplicării unei alte serii de prelucrări, la modulul următor.

În varianta actuală a sistemului am implementat două tipuri de blocuri, ilustrate în Figura 1, acestea fiind suficiente pentru operațiile uzuale într-un flux de date:

– blocuri de procesare – module ce vor genera sau altera textul
– blocuri decizionale – stabilesc drumul care va fi urmat

3.1 Arhitectura sistemului

Sistemul de colectare de corpusuri multilingve este compus din două module distincte: editorul de fluxuri și serviciul de colectare a paginilor web.

EDITORUL DE FLUXURI. În cadrul editorului de fluxuri se realizează schema de funcționare a aplicației, iar editarea diagramei poate fi făcută în mod vizual. De asemenea, pot fi editați parametrii lansării în execuție a proceselor ce vor alcătui sistemul de extragere, iar blocurile pot fi testate în mod individual.

SERVICIUL DE COLECTARE A PAGINILOR WEB este un serviciu Windows care rulează la un interval de timp prestabilit. Parametrii de configurare a serviciului sunt încărcați dintr-un fișier XML. Documentul XML reprezintă schema de funcționare a aplicației creată cu ajutorul editorului de fluxuri. Serviciul de colectare a paginilor web, conform diagramei, lansează în execuție procesele și asigură comunicarea între ele. De asemenea, serviciul dispune de un jurnal detaliat al operațiilor efectuate cu ajutorul căruia pot fi investigate posibilele erori ale proceselor lansate.

3.2 Modul de utilizare

Pentru colectarea de corpus multilingv dintr-un anumit domeniu al web-ului este necesară crearea unei diagrame cu ajutorul aplicației de editare de fluxuri. În această diagramă sunt specificate procesele ce urmează a rula. De asemenea, în diagramă trebuiesc stabiliți parametrii de lansare în execuție a proceselor (interpretorul utilizat, parametrii în linie de comandă etc.) și ce expresii regulate vor fi aplicate intrării/ieșirii proceselor. Pentru algoritmii în care este luată o decizie, în diagramă se va completa și regula de validare a conținutului.

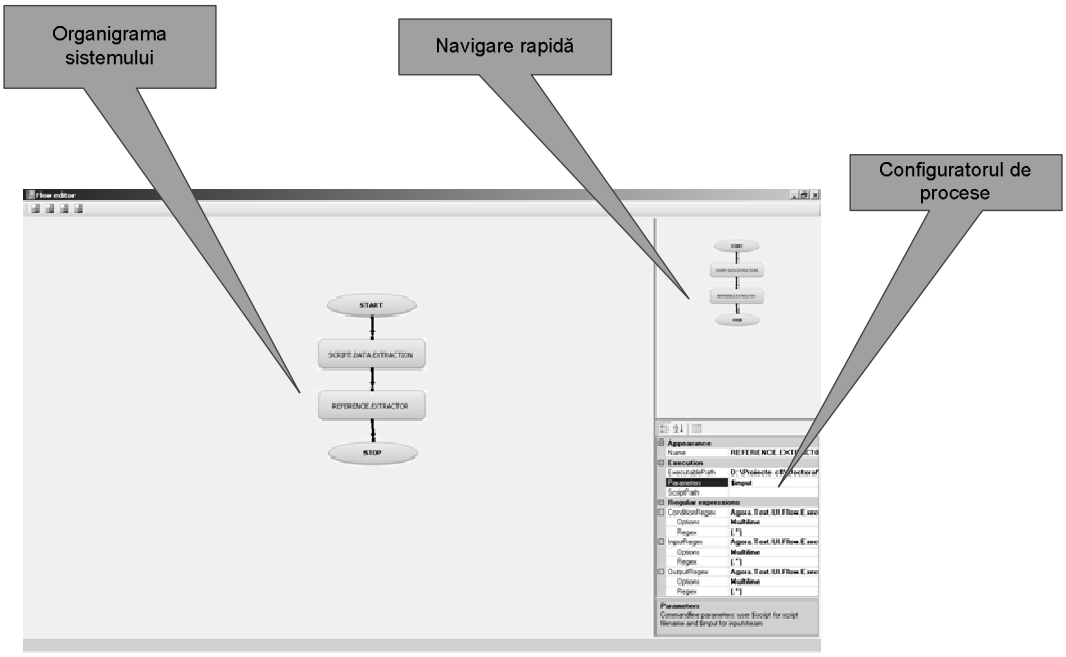


Figura 2: Editorul de diagrame

Fiecare bloc din diagramă reprezintă o unitate de execuţie. Aceasta este caracterizată de calea către numele programului ce urmează a fi lansat, parametrii liniei de comandă (cuvintele cheie $script şi $input vor fi înlocuite cu numele script-ului şi, respectiv, intrarea obţinută la pasul anterior) şi expresiile regulate folosite (pentru intrare, ieşire şi condiţia de validare).

Odată definită, diagrama de flux este salvată într-un fişier XML ce este importat de sistemul propriu-zis de extragere. Acesta execută întregul flux la intervalul de timp stabilit în prealabil. Diagrama poate fi modificată în orice moment, urmând ca aceasta să fie reîncărcată la următoarea pornire a aplicaţiei de colectare de pagini web.

4. Cazuri practice

Prezentăm în continuare două exemple de utilizare a aplicaţiei: (i) pentru constituirea de corpus multilingv din domeniul turismului (*Wikitravel*) şi pentru colectarea de corpus puternic comparabil din domeniul jurnalistic (*Parlamentul European* - secţiunea de ştiri). În ambele cazuri s-au folosit scheme de procesare simple similare cu cea prezentată în Figura 2. Au fost utilizate doar două programe/procese şi nu au fost necesare blocuri decizionale. Cele două procese constau în: (i) generarea unei liste de adrese ce urmează a fi scanate şi (ii) parcurgerea acestora şi prelucrarea rezultatelor.

4.1 Wikitravel

În cazul site-ului Wikitravel s-a urmărit extragerea de corpusuri puternic comparabile din registrul textual al turismului pentru limbile română, engleză şi germană. Cu ajutorul unei expresii regulate au fost extrase adresele paginilor din limba română

disponibile în domeniul „wikitravel.org”. Au fost urmate link-urile ce se regăseau în versiunea pentru limba română a paginii, urmând ca traducerile să fie găsite prin căutarea textului în subdomeniile /de/ sau /en/. Pentru validarea adreselor candidate au fost folosite liste pentru secvențele de adresă permise și pentru secvențele de adresă nepermise, prezentate în tabelul 1, prioritatea pentru validare fiind acordată secvențelor permise.

Tabel 1: Reguli de validare a adreselor

Secvențe de adresă permise	Secvențe de adresă nepermise
wikitravel.org/de/	#
wikitravel.org/en/	mailto
wikitravel.org/ro/	special
	action=
	Special:
	doubleclick
	User:
	Wikitravel:
	Utilizator:
	Image:
	http:
	www.

Legăturile care reprezentau traduceri ale articolului pentru engleză și germană au fost grupate, iar în cazul în care acestea lipseau, au fost înlocuite generic cu o pagină goală. Procesul a fost continuat prin extragerea textului și eliminarea elementelor HTML din date.

Tabel 2: Rezultatele obținute în cazul Wikitravel

română - engleză	română-germană
400.000 cuvinte	100.0 vinte

4.2 Parlamentul European

Pe site-ul Parlamentului European la secțiunea știri se găsesc articole traduse în 22 de limbi. La fel ca în cazul Wikitravel, pornind de la adresa http://www.europarl.europa.eu/news/public/toute_actualite/default/default_ro.htm se parcurg toate referințele găsite pe pagină. Spre deosebire de cazul anterior, un site de știri se actualizează mai des, astfel că datele cantitative prezentate în Tabelul 3 sunt cele de la momentul scrierii acestui articol.

Tabel 3: Corpus obținut

Limbă	Cod	Cuvinte
bulgară	bg	54351
cehă	cs	57480
daneză	da	67856

Limbă	Cod	Cuvinte
italiană	it	80431
lituaniană	lt	80431
letonă	lv	67211

Limbă	Cod	Cuvinte
germană	de	70307
greacă	el	90729
engleză	en	84484
spaniolă	es	91447
estoniană	et	39388
finlandeză	fi	62692
franceză	fr	89171
ungară	hu	59584

Limbă	Cod	Cuvinte
malteză	mt	46041
olandeză	nl	74390
poloneză	pl	56031
portugheză	pt	75557
română	ro	81433
slovacă	sk	67189
slovenă	sv	68794

Sperăm ca în câteva luni, corpusul multilingv EU-News (22 limbi) să devină o resursă publică, de dimensiune corespunzătoare, pentru cercetări şi dezvoltări în domeniul cros şi multilingv (traducere automată, sisteme de regăsire documentară multilingvă, sisteme de întrebare/răspuns cros-lingvistice, etc.).

În ambele exemple prezentate, corpusurile au fost stocate în directoare distincte de fişiere de tip text. Documentele paralele/comparabile au fost numite folosind identificatori unici extensiile acestora denotând codul ISO al limbii documentului (de pildă nume de tipul xxxxxxxxx.bg, xxxxxxxxx.cs, ..., xxxxxxxxx.sv identifică fişierele conţinând documente paralele/comparabile în cele 22 de limbi ale corpusului multilingv). În plus, au fost generate metadate corespunzătoare fiecărui fişier text, metadate ce documentează calea către fişier, adresa de la care a fost extras fişierul, numărul de cuvinte din fişier etc.

5. Concluzii

Rezultatele bune obţinute în urma colectării de corpus multilingv din site-ul Wikitravel şi din secţiunea de ştiri a site-ului Parlamentului European dovedesc utilitatea setului de instrumente propus în această lucrare. Aceste instrumente facilitează interoperabilitatea diverselor aplicaţii scrise în diferite limbaje de programare, fiecare aplicaţie fiind reprezentată ca un proces al unei diagrame de flux. De asemenea, posibilitatea de a reconfigura în orice moment a diagrama de flux permite abordarea unor proiecte de colectare de corpus de complexitate ridicată – procesele, şi regulile de validare pot fi schimbate chiar în cursul colectării documentelor.

Mulţumiri. Activitatea de cercetare descrisă a fost sprijinită de Comisia Europeană prin proiectul ACCURAT (248347/FP7) şi (parţial) de CNCSIS – UEFISCSU prin proiectul PNII – IDEI STAR, 742/19.01.2009

Referinţe bibliografice

Bertoldi, N., Haddow, B., Fouet, J.-B. (2009). *Improved Minimum Error Rate Training in Moses*, în Prague Bulletin of Mathematical Linguistics, Nr. 91, 2009, pp. 7-16.

Ceauşu, A. (2009). *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă*, Bucureşti, România: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Chappell, D. (2009). The Workflow Way - Understanding Windows Workflow Foundation.

Germann, U. (ed.) (2001). *Aligned Hansards of the 36th Parliament of Canada* - Release 2001-1a. http://www.isi.edu/natural-language/download/hansard/ (15.03.2010).

Hoang, H., Koehn, P., Lopez, A. (2009). A Uniform Framework for Phrase-Based, Hierarchical and Syntax-Based Machine Translation, International Workshop on Machine Translation (IWSLT), pp. 152-159.

Koehn, Ph. (2005). *EuroParl: A Parallel Corpus for Statistical Machine Translation*. Machine Translation Summit 2005. Phuket, Thailand. http://www.statmt.org/europarl/(16.03.2010).

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, Ch., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Ch., Zens, R., Dyer, Ch., Bojar, O., Constantin, A., Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Kobayashi, M. and Takeda, K. (2000). „Information retrieval on the web". *ACM Computing Surveys* (ACM Press) **32** (2): 144–173. doi:10.1145/358923.358934

Och, F. J. (2003). *Minimal Error Rate Training in Statistical Machine Translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation*. In ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., Varga, D. (2006). *The JRC-Acquis: A Multilingv Aligned Parallel Corpus with 20+ Languages.* In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 2142-2147, Genoa, Italy, May 2006. ELRA - European Language Ressources Association.

Tiedemann, J. (2009). *News from OPUS - A Collection of Multilingv Parallel Corpora with Tools and Interfaces*. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia.

Tufiș, D., Ion, R., Ceaușu, A., Ștefănescu, D. (2007). *Servicii web lingvistice ale ICIA*. În Ionuț Pistol, Dan Cristea, and Tufiș, D. (eds.), Resurse Lingvistice și Instrumente pentru Prelucrarea Limbii Române, pp. 61-68, Iași, România, dec. 2007. Universitatea "Al.I. Cuza" Iași, Editura Universității "Al.I. Cuza" Iași. ISBN 978-97-3703-297-3.

Tufiș, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2009). Building Language Resources and Traslation Models for Machine Translation Focused on South Slavic and Balkan Languages. In J. Machacova and K. Rohsmann (eds) „Scientific Results of the SEE-ERA.NET Pilot Joint Call". Center for Social Innovation Publisher, Vienna, ISBN 978-3-200-01567-8, pp. 37-48, June 2009.

PARSAREA COMPARATIVĂ A DICȚIONARELOR-TEZAURE ROMÂNEȘTI, FRANCEZE ȘI GERMANE

NECULAI CURTEANU[1], MIHAI ALEX MORUZ[1,2], DIANA TRANDABĂȚ[1]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iași*
[2] *Universitatea „Al. I. Cuza", Facultatea de Informatică, Iași – România;*

{curteanu, mmoruz, dtrandabat}@iit.tuiasi.ro

Rezumat

Lucrarea de față prezintă un studiu comparativ privind modelarea lexico-semantică și parsarea dicționarelor tezaure pentru diferite limbi. Transformarea eficientă a unui tezaur într-o bază de date structurată de arbori lexico-semantici se bazează pe metoda configurațiilor SCD (Segmentare-Coeziune-Dependență), acestea fiind aplicate succesiv pentru identificarea segmentelor lexicografice și extragerea arborelui de sensuri și sub-sensuri din intrare. Modelarea lexico-semantică dezvoltată pentru tipologia intrărilor de dicționar din Dicționarul tezaur al Limbii Române - format *nou* (**DLR**) a fost adaptată și aplicată la parsarea altor patru tezaure: Dicționarul tezaur al Limbii Române - formatul *vechi* (**DAR**), Trésor de la Langue Française (**TLF**), Deutsches Wörterbuch – GRIMM (**DWB**), și Göthe-Wörterbuch (**GWB**). Lucrarea prezintă clasele de markeri de sensuri și hipergraful de dependențe specifice pentru fiecare dicționar, parsarea acestor tezaure, și discuții privind rezultatele obținute și analiza erorilor.

1. Introducere

Parsarea unui dicționar presupune transformarea intrărilor ce conțin text sub formă de glosă, într-un format indexabil. Astfel, fiecare intrare de dicționar este transpusă într-o structură complexă, care conține atât sensurile definite, precum și descrieri detaliate ale formei intrării, cu referire la ortografie, morfologie, fonetică, etimologie, uz etc.

Scopul acestei lucrări este analiza impactului configurațiilor SCD (Segmentation-Cohesion-Dependency) (Curteanu; 2006) la parsarea intrărilor de dicționar pentru patru dicționare tezaure: Dicționarul tezaur al Limbii Române - formatul vechi (**DAR**), Trésor de la Langue Française (**TLF**), Deutsches Wörterbuch – GRIMM (**DWB**), și Göthe-Wörterbuch (**GWB**). Metoda a fost deja aplicată cu succes pentru parsarea Dicționarului tezaur al Limbii Române - formatul nou (**DLR**) (Curteanu et al.; 2008a, 2008b, 2009a), folosind configurații SCD. Prima configurație are rolul de a identifica segmentele lexicografice distincte dintr-o intrare de dicționar. În cadrul strategiei dezvoltate pentru **DLR,** alte două configurații diferite de parsare sunt folosite: una care identifică și extrage, pentru fiecare intrare din dicționar, arborele specific de sensuri (ierarhia sensurilor principale și secundare), și o altă configurație care parsează fiecare nod din arborele de sensuri cu scopul de a clasifica definițiile acelui sens/subsens din dicționar. Spre deosebire de metodele standard de parsare a textului de dicționar, în care toate câmpurile unei intrări de dicționar sunt analizate secvențial, noua metodă reușește detașarea procesului de construire a arborelui de sensuri (a *doua configurație* SCD) de procesul parsării la definițiile sensurilor (a *treia configurație* SCD). Separarea celor

două procese se face în principal prin selectarea *breadth-first* a tuturor marcherilor la sensuri, urmată de analiza *depth-first* a secvenţelor de marcheri pentru fiecare intrare de dicţionar. Amintim că o *primă configuraţie* SCD realizează recunoaşterea segmentelor lexicografice ale intrării.

Spre deosebire de orientările standard în parsarea intrărilor de dicţionar (Neff & Boguraev; 1989), de exemplu sistemul **LexParse** (Hauser & Storrer; 1993), (Kammerer; 2000), (Lemnitzer & Kunze; 2005) sau gramaticile lexicografice (Curteanu & Amihăesei; 2004), (Tufiş; 2001), metoda descrisă în (Curteanu et al., 2008a) *detaşează complet* procesul de construire a arborilor de sensuri de procesul parsării definiţiilor la sensuri. O configuraţie SCD are următoarele componente:

- Un set de clase de marcheri: un marcher reprezintă o graniţă computaţional-lingvistică pentru o categorie lingvistică specifică;
- O ierarhie de tip arbore (sau graf aciclic orientat, sau hipergraf), care stabileşte dependenţele dintre clasele de marcheri;
- Un algoritm de parsare, care execută următorii paşi: recunoaşterea marcherilor, identificarea structurilor textuale dintre doi sau mai mulţi marcheri, şi clasificarea acestor structuri ţinând cont de ierarhia claselor de marcheri. Pentru aplicarea configuraţiilor SCD la parsarea dicţionarelor este necesară o abordare lexico-semantică a claselor de marcheri de sensuri şi definiţii, şi stabilirea unei reprezentări ierarhice a lor.

Lucrarea de faţă prezintă configuraţiile SCD pentru cele patru dicţionare-tezaur considerate:

- *Recunoaşterea segmentelor lexicografice care corespund unei intrări de dicţionar*: Pentru dicţionare complexe / tezaure, fiecare intrare este compusă din mai multe segmente lexicografice, întinderi textuale care partiţionează intrarea de dicţionar în unităţi lexicografice distincte. Astfel, o *primă configuraţie* SCD este dedicată obţinerii segmentelor lexicografice specifice unei intrări.
- *Construcţia arborelui de sensuri al unei intrări*: După recunoaşterea segmentelor lexicografice, acestea sunt clasificate în funcţie de structura lor lingvistică şi semantică. Cel mai important segment este cel al *descrierii sensurilor*, asupra căruia se aplică o *a doua configuraţie* SCD pentru extragerea arborelui de sensuri al intrării. Pentru **DLR**, arborele de sensuri este obţinut cu o precizie de 91.18% (Curteanu et al.; 2008a, 2009a).
- *Clasele de marcheri* folosite pentru identificarea arborelui de sensuri al unei intrări din **DLR** sunt, pentru *sensuri principale*: *majuscule* (**A.**, **B.,** etc.); *cifre romane (***I.**, **II.**, etc.); *cifre arabe* (**1.**, **2.**, etc.). Pentru *sensuri secundare*: *rombul plin* ♦; *rombul gol* ◊; *enumerarea literală* cu litere latine mici **a)**, **b)**, **c)**,…
- *Segmentarea definiţiilor*: A *treia configuraţie* SCD foloseşte un set specific de clase de marcheri pentru segmentarea la definiţii atomice în fiecare sens al intrării şi stabilirea dependenţelor dintre sensuri (vezi (Curteanu et al.; 2009a) pentru o clasificare a definiţiilor atomice). Analiza intrărilor **DLR** a dus la următoarele tipuri de definiţii: **(i)** *MorfDefs* – definiţii morfologice; **(ii)** *RegDefs* – definiţii scrise cu font regular; **(iii)** *BoldDefs* – definiţii scrise cu bold; **(iv)** *ItalDefs* – definiţii scrise cu italic; **(v)** *SpecDefs* – definiţii ce conţin specificaţii; **(vi)** *SpSpecDefs* - definiţii scrise cu litere spaţiate, ce conţin anumite specificaţii; **(vii)** *DefExems* – exemple la definiţii, cu rolul de a întregi înţelesurile unei definiţii. Tipurile de definiţii propuse aici primesc roluri funcţionale specifice în

descrierea sensurilor principale, secundare, sau de granularitate semantică mai fină (Curteanu et al.; 2009a).

2. Parsarea Dicționarului tezaur al limbii române– formatul vechi (DAR)

Structura principalelor segmente lexicografice ale unei intrări **DAR** este următoarea:

I. Segmentul *traducerii franceze*, *FreSeg*, conține lema și structura principalelor sensuri ale cuvântului, traduse în limba franceză. Comparația dintre traducerea structurii sensurilor lemei și descrierea completă a intrării **DAR** pentru limba română revine la aplicarea operației de *subsumare* între arborele de sensuri al cuvântului românesc și arborele de sensuri al traducerii lui în limba franceză. Există multiple situații în această comparare, inclusiv aceea în care perechea <$cuvânt_{Rom}$, mot_{Fre}> nu poate fi determinată: de exemplu <$mămăligă_{Rom}$, $?_{Fre}$>.

II. Segmentul *descrierii generale* a lemei din limba română (*RomSeg*). *RomSeg* conține, uneori pe mai multe *paragrafe*, informații morfo-sintactice, semantice, de utilizare etc. ale intrării. *RomSeg* începe, de obicei, cu lema în format italic (uneori lema se află pe primul rând al primului paragraf din *RomSeg*).

III. Al treilea segment al unei intrări din **DAR**, *SenseSeg*, este chiar *descrierea lexical-semantică* a cuvântului-lemă. *SenseSeg* constituie principalul obiectiv al analizei lexico-semantice și al programului de parsare la arborele de sensuri în **DAR**.

IV. Al patrulea segment, *NestSeg*, al unei intrări în **DAR** este format din unul sau mai multe „*cuiburi*", întinderi de text lexical-semantic în care sunt descrise variante morfologico-sintactice, fonologice, regionalisme etc. ale intrării, uneori cu descrierea subsidiară a sensurilor specifice. Structura segmentului-*cuib* din **DAR** este similară cu cea a unei intrări generale din **DAR**. Construcția posibil recursivă a intrărilor **DAR** rezultă din necesitatea parsării la sensuri și într-un segment-*cuib*.

V. Al cincilea segment al intrărilor din **DAR**, denotat *EtymSeg*, face precizări asupra etimologiei lemei și este introdus printr-o *liniuță-de-etimologie* (cratimă-mare „**–**").

Din cele cinci segmente ale unei intrări **DAR**, singurele obligatorii sunt *FreSeg* și *SenseSeg*. Celelalte trei segmente apar opțional în descrierea intrărilor, în funcție de necesitățile și specificul fiecărei leme.

2.1 Clase de marcheri la sensuri

În ordinea priorităților din ierarhia claselor de marcheri la sensuri **DAR**, avem:

1. Litere latine majuscule (*LatCapLett_Enum*): A., B., C., … 2. Cifre romane majuscule (*LatCapNumb_Enum*): I., II., … 3. Cifre arabe (*ArabNumb_Enum*): 1^0., 2^0., …

În mod similar cu clasificarea sensurilor din **DLR**, vom considera că marcherii acestor clase introduc *sensuri principale* și în **DAR**.

4. Pentru a introduce *sensuri secundare*, de obicei pentru definițiile *obligatorii* și/sau *opționale* (Curteanu et al.; 2009a), **DAR** folosește marcheri de sens care introduc și în **DLR** definițiile de tip *MorfDef*, *RegDef*, *BoldDef*, *ItalDef*, *SpecDef*, *SpSpecDef*, și *DefExem*, dar și câțiva *marcheri specifici* **DAR**: ||, |, #, †. **DAR** definește primii doi

marcheri astfel: „Liniuţele verticale despart diferitele subsensuri, şi anume liniuţa simplă | pe cele mai puţin deosebite, iar liniuţa dublă || pe cele mai marcante. Uneori, la verbe, aceste liniuţe despart pe *transitive*, de *intransitive* şi de *reflexive*". Marcherul # delimitează „idiotismele şi locuţiunile proverbiale", iar † introduce sensurile sau expresiile scoase din uzul limbii (Curteanu et al.; 2009b).

5. În funcţie de nivelul din arborele de sensuri al descrierii lexical-semantice, situându-se fie pe sensuri principale fie secundare, **DAR** foloseşte şi *enumerarea literală* pe *două niveluri*: (5.a) Litere latine minuscule (*LatSmaLett_Enum*): a.), b.), … (5.b) Sub o enumerare literală *LatSmaLett_Enum* se poate face o a doua enumerare, cu litere greceşti minuscule (*GreSmaLett_Enum*): α.), β.), γ.), …

2.2 Ierarhia claselor de marcheri pentru dependenţele la sensuri în DAR

Introducem următoarele ierarhii şi subierarhii pentru clasele de marcheri la sensuri şi definiţii din **DAR**:

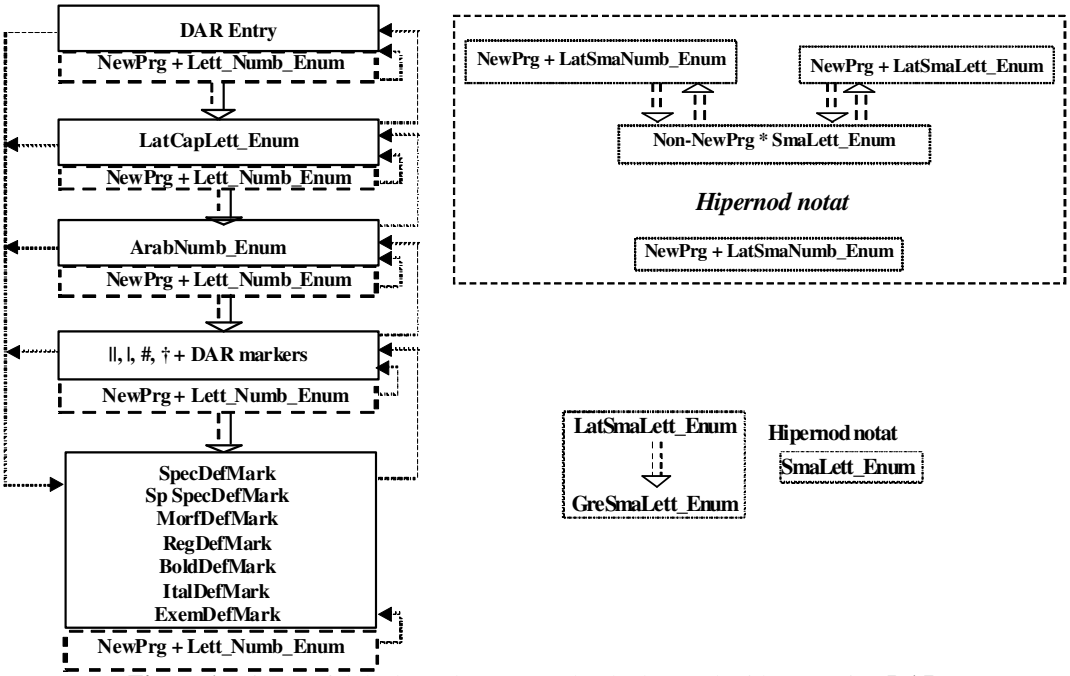


Figura 1: Hipergraful de dependenţe între clasele de marcheri la sensuri în **DAR**

Hipergraful de dependenţe la clase de marcheri **DAR** (Curteanu et. al.; 2009b) este mult mai complex (enumerări la paragrafe noi şi enumerări literale) în comparaţie cu hipergraful pentru **DLR** (Curteanu; 2008a, 2009a).

2.3 Probleme speciale şi rezultate ale parsării DAR

Principala *dificultate* a parsării intrărilor din **DAR** este folosirea marcherului de sens *Paragraf_Nou* (*NewPrg*), aspect care *nu* apare în **DLR**. Marcherul *NewPrg* este costisitor de recunoscut (din punct de vedere al timpului de execuţie) deoarece, pentru acest marcher, a trebuit să introducem un *nivel implicit* de *enumerare* dacă şi numai dacă *NewPrg* ***nu*** este însoţit şi de un alt tip de marcher la sensuri **DAR** (în special

marcheri de sensuri principale). Enumerarea implicită generată de marcherul *NewPrg* se face cu *cifre latine minuscule* (*LatSmaNumb_Enum*): *i, ii, iii, iv, v, vi, vii* etc.

A doua problemă dificilă care apare la parsarea **DAR** este procedeul de rafinare a marcherului *NewPrg* prin enumerare literală cu latine minuscule *LatSmaLett_Enum*, care pot, la rândul lor, să fie rafinate printr-o enumerare implicită de tipul *NewPrg*. Această problemă a fost rezolvată atribuind marcherilor de tip *NewPrg,* și enumerării cu latine minuscule, niveluri de sens care variază în funcție de context.

Un exemplu remarcabil este intrarea **CAL** din **DAR**, în care aspectul ce trebuie subliniat este că *NewPrg*, enumerat cu cifre latine minuscule *LatSmaNumb_Enum* (pe primul subnivel de dependență), apelează enumerarea literală cu *NewPrg* (la paragraf nou) – al doilea subnivel de dependență, care la rândul ei apelează *NewPrg* enumerat (al treilea subnivel). De pe acest nivel de *NewPrg* este apelată din nou enumerarea literală la *Non-NewPrg* (al patrulea subnivel de dependență). Toate cele *cinci* niveluri componente de dependență se atașează formal nivelului de sens reprezentat de marcherul 1^0. Aceste apelări reciproce și succesive se bazează pe hipergraful de dependențe din Fig. 1.

Parsarea **DAR** este descrisă în detaliu în (Curteanu et al.; 2009b). Au fost parsate un număr de 37 intrări **DAR**, de dimensiuni mari și medii. Evaluarea manuală a parsării este foarte promițătoare având în vedere problemele ridicate de parsarea **DAR**, semnificativ mai dificile decât la **DLR**. În lipsa unui corpus gold pentru parsarea intrărilor **DAR**, nu s-a putut face o evaluare automată a parserului.

3. Parsarea TLF cu configurații SCD

3.1 Segmente lexicografice în TLF

Dicționarul-tezaur **TLF** prezintă și asemănări cu **DLR** dar și aspecte distincte interesante. Ca observație generală, **TLF** este unul din cel mai bine structurate dicționare-tezaur dintre cele analizate. Configurația SCD-*Config*1 realizează diviziunea la *segmente lexicografice* a intrărilor din **TLF**, structura acestor segmente fiind relativ simplă. O intrare **TLF** începe chiar cu segmentul de *descriere a sensurilor*. Acest segment este urmat, opțional, de o serie de segmente *finale* introduse de etichete **TLF** bine definite. Aceste segmente pot fi văzute și ca un singur segment lexicografic, format din componente opționale, specifice fiecărei etichete. Calupul de segmente finale (sau segmentul final) este introdus de etichete după modelul ce urmează:

REM. 1. … **2.** … **3.** …
PRONONC. ET ORTH. – … Eng.: …
ÉTYMOL. ET HIST. I. ... **1. a)** … **b)** … **2.** … **3.** … **II**. …
STAT. Fréq. abs. littér.: … **Fréq. rel. littér.:** …
DÉR. 1. … **2.** … **3. a)** … **b)** … … **Rem. a)** … **b)** …
BBG. …

Există intrări **TLF** în care aceste segmente componente nu încep la paragraf nou *NewPrg* ci formează împreună un paragraf compact, segmentele componente fiind delimitate de etichetele de mai sus (în unele cazuri cu *majuscule-bold*, în alte cazuri cu *minuscule-bold*, sau amestecat). După cum se observă din șablonul de mai sus, unele

segmente finale conțin descrieri particulare de sensuri, perfect similare cu descrierile din segmentul propriu-zis de sensuri.

3.2 Clase de marcheri TLF și hipergraful de dependență

Marcherii de sensuri din **TLF** sunt oarecum asemănători cu cei din **DLR** dar există și diferențe semnificative, ce impun până la urmă specificul de construcție și de reprezentare al celor două dicționare-tezaure. În comparație cu **DLR**, în **TLF** avem:

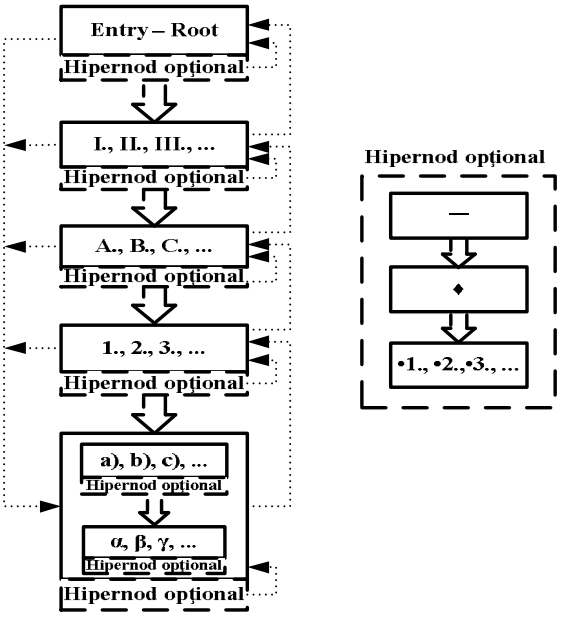


Figura 2: Hipergrafurile de dependențe la clasele de marcheri **TLF**

(1) *Marcheri noi*: **(1.1)** Marcherul „**–**” (*liniuța-de-moștenire*) introduce un *subsens moștenit*. Pentru marcherul „**–**” de moștenire a unui sens avem *două* situații importante în intrările **TLF**: **(a)** Atunci când „**–**” apare *după* un alt marcher de sens din **TLF**, rolul „**–**” este cel de moștenire a sensului de pe nodul-părinte de sens (*regent* sau *nu*) din arborele de sensuri al intrării; **(b)** Când „**–**” este la paragraf nou, deci *NewPrg* + „**–**”, liniuța-de-moștenire „**–**” are rol de *subsens intermediar*, similar cu rolul *NewPrg* din tezaurul **DAR**. În acest caz particular, „**–**” introduce un subsens al marcherului / sensului-părinte *regent*, *direct* dependent de (sub)sensul regent. **(1.2)** Marcherul •1., •2., … (*bulină-roșie indexată*) definește un ***concept nou***, ***specific*** **TLF**: *DefExems indexate* pe întreaga intrare **TLF** (*IdxDefExem*). **(1.3)** *Enumerarea literală dublă*: enumerarea *LatSmaLett_Enum* (latine mici **a)**, **b)**, …) este rafinată prin *GreSmaLett_Enum* (litere grecești mici **α)**, **β)**, …).

(2) În **TLF** întâlnim ca marcher de sens secundar numai *rombul-plin* „♦”, nu și rombul-gol, prezent în **DLR**.
(3) În comparație cu **DLR**, în **TLF** apare o inversare a dependențelor la clasele de marcheri pentru sensurile principale, reflectată în hipergraful de dependențe din Fig. 2.

3.3 Rezultate de parsare la sensuri pentru marcheri expliciți în TLF

Pentru **TLF** au fost parsate un număr de 31 intrări, cu dimensiuni medii și mari (în medie, 5 pagini format A4). Rezultatele parsării **TLF** au fost evaluate manual, estimarea preciziei parsării fiind foarte bună. Principalul motiv al acestor rezultate este elaborarea consecventă a intrărilor **TLF**, structura lor regulată, apariția foarte rară a excepțiilor de la ierarhia de marcheri propusă. În lipsa unui corpus gold pentru parsarea intrărilor **TLF**, nu s-a putut face o evaluare automată (precisă) a parserului.

4. Segmente lexicografice și clase de marcheri la sensuri în DWB și GWB

4.1 Moduri de recunoaștere a segmentelor lexicografice din DWB

Intrările **DWB** sunt constituite din mai multe *segmente lexicografice*. Aceste segmente au o structură neuniformă și neunitară. *Caracteristic* pentru tezaurele germane **DWB** și **GWB** este structura segmentelor lexicografice, formate dintr-un prim subsegment (opțional) pe care l-am numit *sensul-rădăcină* al segmentului, și din al doilea subsegment, *corpul propriu-zis* al segmentului lexicografic. În timp ce sensurile incluse în *corpul* fiecărui segment lexicografic sunt descrise cu marcheri expliciți, relativ ușor de recunoscut, parsarea completă a segmentelor **DWB** nu este o sarcină ușoară. În **DWB**, segmentele lexicografice sunt introduse în *trei* moduri:

(A) După *sensul-rădăcină* (*definiția-rădăcină*) al unei intrări **DWB**, sau după *sensul-rădăcină* al unui segment lexicografic, sunt poziționate un număr de cuvinte-rezervate, scrise (de obicei) *spațiat* și cu *font-italic*. Aceste cuvinte-rezervate (sau *listă* de cuvinte rezervate) se constituie în *eticheta segmentului lexicografic* a cărui întindere **urmează** segmentului curent din **DWB**. Exemple de asemenea cuvinte-cheie ce delimitează începutul unui segment lexicografic: „*Form, ausbildung und ursprung*", „*Formen*", „*Ableitungen*", „*Verwandtschaft*", „*Verwandtschaft und form*", „*Formelles und etymologisches*", „*Gebrauch*", „*Herkunft*", „*Grammatisches*", etc., sau, pentru segmentul de descriere a sensurilor din **DWB**, „*Bedeutung und gebrauch*" (sau numai „*Bedeutung*"). În exemple, etichetele de segmente lexicografice au fondul 25% gri:

GRUND, *m.*, *dialektisch auch f. gemeingerm. wort; fraglich ist das geschlecht von got.* *grundus *in* grunduwaddjus, *vgl.* afgrundiþa; *sonst meist masc.: ahd.* grunt, crunt; *mhd.* grunt; *as. … … meist f., selten m.; … …* grunte *f., lett.* grunts *m.,* grunte *f., poln. russ. slov. nlaus.* grunt *m. f o r m u n d h e r k u n f t .*

1) *für das verständnis der vorgeschichte des wortes ist die z w i e g e s c h l e c h t i g k e i t ...*

H. V. SACHSENHEIM *spiegel* 177, 30;

die neuen grundt zu der kirchen *zimm. chron.*[2] 2, 539, 36; du findest noch vil gar alter meür und grunt und thürn SIGMUND MEISTERLIN *in städtechron.* 3, 51, 14. … … … drey starcke grund 6, 290. *b e d e u t u n g . die bedeutungsgeschichte des wortes läszt sich schwer aufbauen, … … doch erschöpft diese unterscheidung einer mehr räumlichen und mehr flächenhaften vorstellung die sache nicht.*

I. grund *bezeichnet die feste untere begrenzung eines dinges.*

A. *grund von gewässern; seit ältester zeit belegbar: profundum* (*sc. mare*) crunt *ahd. gl.* 1, 232, 18; latid thea odra (*fisch*) eft an gr. faran *Hel.* 2633.

1) *am häufigsten vom meer* (*in übereinstimmung mit dem anord. gebrauch*): ...

(B) *Al doilea* mod de specificare a segmentelor lexicografice din **DWB** este cel prin care *imediat după* marcherii de sensuri principale sunt specificate cuvintele-cheie ce reprezintă *eticheta* segmentului lexicografic care urmează.

GEBEN, *dare.*
I. *Formen, ableitungen, verwandtschaft.*
1) *es ist ein allgemein, aber ausschlieszlich germanisches wort: goth.* giban (*praet.* gaf), *ahd. ...*
II. *Bedeutung und gebrauch.*
1) geben *und* nehmen, *die beiden sich ergänzenden gegenstücke, verdienen die erste…*

(C) *Al treilea* mod de specificare a descrierii lexical-semantice a unei intrări în **DWB** (deci şi a segmentului de descriere a sensurilor), care este şi cel mai frecvent, este acela în care nu este folosită nicio etichetă pentru începutul marcherilor de sensuri. Implicit, după sensul rădăcină (care poate fi redus doar la traducerea latină a cuvântului lemă din **DWB**), urmează segmentul de descriere a sensurilor (care este şi unic), *fără eticheta explicită* „*Bedeutung*", doar prin prezenţa expresă a marcherilor de sensuri, a definiţiilor atomice, a exemplelor la definiţii şi a siglelor pentru aceste exemple.

4.2 Clasele de marcheri şi dependenţele la sensuri în DWB şi GWB

Marcherii la sensuri sunt cei obişnuiţi din dicţionarele-tezaur, cu menţiunea că rafinarea sensurilor prin *enumerare literală* se face în **DWB** pe trei niveluri, *LatSmaLett_Enum*, *GreSmaLett_Enum*, dar şi *Gre2SmaLett_Enum* (αα), ββ), …), iar în **GWB** avem doar primele două. **GWB** conţine şi marcherul *liniuţa-de-moştenire* (ca în **TLF**). Hipergrafurile de dependenţe la sensuri în **DWB** şi **GWB** sunt în Fig.3. şi Fig.4.:

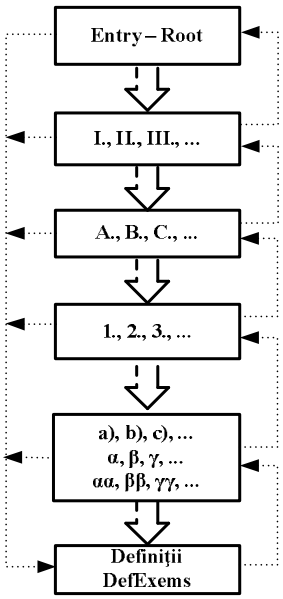


Figura 3: Hipergraful de dependenţe a claselor de marcheri **DWB**

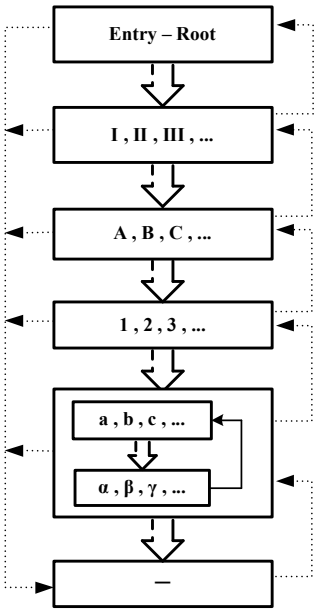


Figura 4: Hipergraful de dependenţe a claselor de marcheri **GWB**

4.3 Rezultate de parsare la segmente lexicografice în DWB

Au fost parsate un număr de 17 intrări de mari dimensiuni din **DWB**, cu diverse particularităţi de redactare. La evaluarea manuală ne-am concentrat asupra rezultatelor de parsare a primelor două configuraţii SCD. Parsarea segmentului de descriere de sensuri se face cu mare precizie, însă separarea segmentelor lexicografice este mai puţin precisă, în mare măsură din cauza structurii complexe a segmentelor lexicografice în

DWB și a variației foarte mari a modului în care acestea sunt marcate. În lipsa unui corpus gold, nu s-a putut face o evaluare automată precisă a parserului.

5. Discuții

Este important să punem în evidență, comparativ lingvistic, rolul funcțional al *structurilor* lexico-semantice, *de sens*, cu cel al *marcherilor* la aceste structuri în dicționarele-tezaure analizate. Noile denumiri propuse au scopul de a stabili o mai strânsă legătură între sensuri și marcherii tipografici / lexicografici prin care sunt identificate și relaționate aceste sensuri.

MorfDef ⇒ *MorfDef*; *RegDef* ⇒ *GlossDef; BoldDef* ⇒ *IdiomDef*; *ItalDef* ⇒ *CollocDef*; *DefExem* ⇒ *DefExem*; *SpecDef* ⇒ *SpecDef*; *SpSpecDef* ⇒ *SpSpecDef*; *IdxDefExem* (concept nou: *DefExem* indexat).
Câteva argumente care susțin propunerile de mai sus: **(1)** Perechea *RegDefs–DefExems* este reprezentată în **DLR** (ca și în **DAR** și **TLF**) prin perechea de marcheri-fonturi *Regular–Italic*, în timp ce aceeași pereche de noțiuni este denotată în tezaurele **DWB** și **GWB** prin perechea ***inversă*** de marcheri-fonturi: *Italic–Regular*. Deci concepte similare sau identice pot fi codificate, în tezaure diferite, prin marcheri tipografici distincți. **(2)** Anumite noțiuni lexico-semantice pot fi particulare unui anumit dicționar-tezaur; așa cum este *DefExem Indexat* (codificat *IdxDefExem*), introdus în **TLF** printr-o bulină roșie indexată. *IdxDefExem* este atașat unei anumite definiții de sens, dar acest *DefExem* particular poate fi referit în mod specific pentru intrarea **TLF** în care este definit.

Experiența acestei lucrări a demonstrat încă o dată generalitatea și eficiența strategiei de parsare cu *configurații* SCD, în acest caz, a textelor de dicționare-tezaure complexe, fiecare cu multe caracteristici specifice. Principala realizare, tradusă în randamentul foarte bun al programelor de parsare, este aceea că analiza se face în primul rând pe secvențele de marcheri extrași din textul intrării, iar procesele de obținere a arborelui de sensuri și de parsare la sensuri / definiții atomice pot fi complet separate (Curteanu et al.; 2008a, 2009a).

Analiza comparativ-lingvistică a celor *patru* dicționare-tezaure a mai relevat faptul că fiecare dicționar are caracteristici particulare și necesită o modelare lingvistică și lexico-semantică atentă. Cu toate acestea, multe concepte referitoare la sensuri, definițiile sensurilor, marcherii de sensuri, exemplele la definiții, referirea sensurilor și siglele pot fi similare, adaptabile și transferabile între configurațiile SCD utilizate, depinzând și de poziția și / sau nivelul la care se situează acestea în cadrul unui arbore de sensuri al unei intrări.

În urma analizării rezultatelor parsării intrărilor din cele *patru* dicționare, am determinat că metoda de parsare propusă, prin folosirea a trei configurații SCD pentru parsarea intrărilor de dicționare-tezaur, este eficientă, robustă și portabilă, putând fi aplicată cu succes peste tezaure (semnificativ) diferite, cu rezultate foarte bune comparativ cu metodele tradiționale de parsare (standard Dictionary Entry Parsing).

Mulțumiri. Rezultatele modelării lexical-semantice și parsării dicționarelor-tezaure **DLR** și **DAR** din această lucrare au fost obținute în cadrul cercetărilor la grantul eDTLR – PNCDI 2, No. 91_013/18.09.2007.

Referințe bibliografice

Comitetul de revizie a DLR (1952). Reguli de codificare pentru DLR. Institutul de Filologie „Iorgu Iordan", Academia Română, București.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In Proceedings of the 4th International IEEE Conference SpeD 2007.

Curteanu, N., E. Amihăesei (2004). *Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries*. ECIT-2004 Conference, Iasi, Romania.

Curteanu, N. (2006). *Local and Global Parsing with Functional* (F)X-*bar Theory and SCD Linguistic Strategy*. (I.+II.), Computer Science Journal of Moldova, Academy of Science of Moldova, Vol. 14 no. 1 (40):74-102; no. 2 (41):155-182.

Curteanu, N., Trandabăț, D., Moruz, A. M. (2008a). *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*, Proceedings of CogAlex Workshop, COLING 2008, pp. 55-63, ISBN 978-1-905593-56-9.

Curteanu, N., Trandabăț, D., Moruz, A., Bolea, C., Husarciuc, M. (2008b). *Parsarea dicționarului eDTLR cu configurații SCD la arborele de sensuri și definiții.* Raport de cercetare la Grantul PNCDI 2, Nr. 91_013/18.09.2007, Faza 2008.

Curteanu, N., Moruz, A., Trandabăț, D. Bolea, C., Spătaru, M., Husarciuc, M. (2009a). *Parsare arborilor de sensuri și segmentarea la definiții în dicționarul tezaur* (Ed. D. Trandabăț, D. Cristea, D. Tufiș) Resurse lingvistice și instrumente pentru prelucrarea limbii române, ConsILR-2008, Editura Univ. „Al. I Cuza" Iași, p. 65-74.

Curteanu, N., Trandabăț, D., Moruz, A., Bolea, C. (2009b). *Parsarea dicționarelor eDA și eDLR cu configurații SCD la arborele de sensuri și definiții.* Raport de cercetare la Grantul PNCDI 2, Nr. 91_013/18.09.2007, Faza 2009.

Erjavec, T, Evans, R., Ide, N., Kilgariff, A. (2000). The CONCEDE Model for Lexical Databases. Research Report on TEI-CONCEDE LDB Project, Univ. of Ljubljana, Slovenia.

Hauser, R., Storrer, A. (1993). *Dictionary Entry Parsing Using the LexParse System*. Lexikographica 9 (1993), 174-219.

Kammerer, M. (2000). *Wöterbuchparsing Grundsätzliche Überlegungen und ein Kurzbericht über praktische Erfahrungen*, http://www.matthias-kammerer.de/content/WBParsing.pdf

Lemnitzer, L., Kunze, C. (2005). *Dictionary Entry Parsing*, ESSLLI 2005.

Neff, M., Boguraev, B. (1989). *Dictionaries, Dictionary Grammars and Dictionary Entry Parsing*, Proc. of the 27th ACL Vancouver, British Columbia, Canada Pages: 91 – 101.

Tufiș, D. (2001). From Machine Readable Dictionaries to Lexical Databases, RACAI, Romanian Academy, Bucharest, Romania.

XCES TEI Standard, Variant P5 (2007): http://www.tei-c.org/Guidelines/P5/

REALIZAREA UNUI TREEBANK ROMÂNESC

CENEL-AUGUSTO PEREZ

Universitatea „Al.I.Cuza", Facultatea de Litere, Iaşi – România

Universitatea „Al.I.Cuza", Facultatea de Informatică, Iaşi – România;

cperez@info.uaic.ro

Rezumat

În lucrarea de faţă vom inventaria rezultatele achiziţionării corpusului de sintaxă a limbii române, punând accent atât pe problemele întâmpinate pe parcursul achiziţionării acestui corpus (se vor menţiona şi câţiva paşi ai creării corpusului), cât şi pe soluţionările acestor probleme.

1. Introducere

Cercetătorii au ajuns la un consens, şi anume, că un progres semnificant şi rapid poate exista atât în înţelegerea textului cât şi în înţelegerea limbii vorbite prin investigarea acelor fenomene lingvistice ce apar foarte frecvent în fapte de limbă întâlnite în mod natural, fără constrângeri, şi prin încercarea de a extrage automat informaţii despre limbă din corpusuri[1] foarte mari. Astfel de corpusuri încep să servească drept unelte importante de cercetare pentru oamenii de ştiinţă din domeniile procesării limbajului natural, a recunoaşterii vorbirii şi a sistemelor de înţelegere a limbajului vorbit, precum şi în lingvistica teoretică.

Între resursele utilizate pentru studierea sintaxei limbilor naturale o componentă importantă sunt treebank-urile. Un treebank este un corpus adnotat la sintaxă într-un formalism general acceptat, ca de exemplu, gramatici de dependenţă, bazate pe relaţii directe existente între cuvinte (Mel'cuk, 1987), adică „relaţii de la cuvânt la cuvânt între cuvinte individuale" (Hudson, 1998), unde toate cuvintele dintr-o frază, cu excepţia unuia (care se numeşte rădăcina frazei) depind de alte cuvinte. Treebank-uri există pentru limbi precum: chineza, ceha, engleza, franceza, germana, italiana, japoneza, poloneza, portugheza, spaniola, turca etc. Cele mai cunoscute treebank-uri actuale sunt pentru limba engleză (Penn Treebank – construit la Universitatea din Pennsylvania, Philadelphia[2]) şi pentru limba cehă (Prague Dependency Treebank – construit la Universitatea Charles din Praga[3]). În dezvoltarea unui parser sintactic pentru limba română (Seretan et al., 2010) s-a utilizat un treebank relativ limitat atât în ceea ce priveşte mărimea lui, cât şi în ceea ce priveşte complexitatea structurii sintactice (adică nu conţine propoziţii subordonate, iar mărimea medie a propoziţiilor este doar de 9 cuvinte). În aceeaşi situaţie se afla şi treebank-ul lui Călăcean şi Nivre (2000), unde întâlnim doar texte din articole de ziare (în special subiecte politice şi administrative) cu

[1] Corpusuri (pl.) – corpus (sg.) = Un corpus este o colecţie de fapte de limbă, sub formă de text electronic, selectate după criterii externe pentru a reprezenta, pe cât posibil, o limbă sau o varietate de limbă ca fiind o sursă de date pentru cercetare lingvistică. (Sinclair 2004)

[2] http://www.cis.upenn.edu/~treebank/

[3] http://ufal.mff.cuni.cz/pdt/

o medie de 8,94 token-uri pe propoziţie. În prezenta lucrare descriem aspecte legate de crearea unei corpus treebank (mai complex) pentru limba română.

În secţiunea a doua vom prezenta succint paşii creării corpusului, iar în secţiunile 3 şi 4 sunt detaliate câteva dintre problemele (însoţite, bineînţeles, de rezolvări) tehnice, respectiv gramaticale, întâlnite în procesul de achiziţionare a corpusului. Ultima secţiune o constituie concluziile unde se vor face o inventariere şi o evaluare a rezultatelor.

2. Crearea corpusului treebank

Crearea unui corpus treebank presupune mai mulţi paşi. Primul pas al acestei construcţii este achiziţionarea surselor lexicale. Am ales texte ce reflectă orientarea spre o selecţie a situaţilor sintactice tipice, din care putem să ne edificăm asupra esenţei combinaţiilor sintactice de diferite tipuri şi pe care le putem utiliza ca modele în diverse analize sintactice mai mult sau mai puţin complicate. S-a urmărit ca exemplele să provină din diverse genuri literare şi să surprindă particularităţile sintactice ale variatelor tehnici de creaţie care au caracterizat o epocă sau alta, ori, mai exact, care au caracterizat arta literară a diferiţilor scriitori. Multe din textele din care e constituit acest corpus sintactic (împreună cu o parte din textele FrameNet-ului englezesc, traduse în limba română) sunt preluate din manualele şi cursuri practice de limba română, din scriitori consacraţi pentru a asigura acurateţea gramaticală şi diversitatea sintactică.

Astfel s-a făcut o selecţie a unor texte, de regulă propoziţii şi fraze cu o întindere între 3 şi 144 cuvinte din operele unor autori precum: Mihai Eminescu, Ion Creangă, Octavian Goga, Liviu Rebreanu, Marin Preda, Mihail Sadoveanu, Barbu Ştefănescu Delavrancea, Calistrat Hogaş, George Coşbuc, Camil Petrescu şi alţii. Aceste texte, împreună cu doar câteva citate din ziare şi almanahuri, au fost fie tehnoredactate de către mine şi transformate în format electronic (format *.doc*) cu ajutorul editorului Microsoft Word din cadrul pachetului de servicii Microsoft Office, fie scanate şi transformate în text electronic prin intermediul unui program OCR (Optical Character Recognition[4]), cu efectuarea corectărilor de rigoare, deoarece operaţia de OCR-izare produce adesea greşeli de interpretare. Astfel, o serie de texte au fost făcute disponibile în format electronic.

Al doilea pas în crearea corpusului îl constituie marcarea automată a informaţiilor de natură morfologică asupra corpusului cu ajutorul webservice-ului de etichetare morfo-sintactică RACAI[5]. De exemplu, luăm fraza: *Rolul covărşitor a lui Eminescu în dezvoltarea literaturii şi a limbii noastre literare a determinat crearea, începînd din acest an, a unei catedre speciale „Mihai Eminescu" la Universitatea din Bucureşti.* Această frază, după procesarea textuală a webservice-ului va arata astfel:

Rolul|rol|NSRY|Ncmsry covărşitor|covărşitor|ASN|Afpms-n a|al|TS|Tsfs lui|lui|TSO|Tf-so Eminescu|Eminescu|NP|Np în|în|S|Spsa dezvoltarea|dezvoltare|NSRY|Ncfsry

[4] O aplicaţie pentru calculator care transformă imagini scanate de texte în şiruri de coduri de caractere, ce vor putea fi astfel căutate, indexate şi prelucrate.
[5] http://www.racai.ro/webservices/TextProcessing.aspx

literaturii|literatură|NSOY|Ncfsoy şi|şi|CR|Crssp a|al|TS|Tsfs limbii|limbă|NSOY|Ncfsoy noastre|meu|PSS|Ds1fsop literare|literar|ASON|Afpfson a|avea|VA3S|Va--3s determinat|determina|VPSM|Vmp--sm creaerea|creare|NSRY|Ncfsry ,|,|COMMA|COMMA începînd|începînd|VG|Vmg din|din|S|Spsa acest|acest|DMSR|Dd3msr---e an|an|NSN|Ncms-n ,|,|COMMA|COMMA a|al|TS|Tsfs unei|un|TSO|Tifso catedre|catedră|NSON|Ncfson speciale|special|ASON|Afpfson „|„|DBLQ|DBLQ Mihai_Eminescu|Mihai_Eminescu|NP|Np „|”|DBLQ|DBLQ la|la|S|Spsa Universitatea|universitate|NSRY|Ncfsry din|din|S|Spsa Bucureşti|Bucureşti|NP|Np .|.|PERIOD|PERIOD

Transformat în xml, acest text va arăta astfel:

```
<?xml version='1.0' encoding='UTF-8'?>
<treebank id="Augusto">
 <sentence id="1" parser="" user="Augusto" date="2010-30-19">
  <word id="1" form="Rolul" lemma="rol" postag="Ncmsry" head="0" deprel=""/>
  <word id="2" form="covărşitor" lemma="covărşitor" postag="Afpms-n" head="0" deprel=""/>
  <word id="3" form="a" lemma="al" postag="Tsfs" head="0" deprel="det."/>
  <word id="4" form="lui" lemma="lui" postag="Tf-so" head="0" deprel=""/>
  <word id="0" form="Eminescu" lemma="Eminescu" postag="Np" head="0" deprel=""/>
  <word id="6" form="în" lemma="în" postag="Spsa" head="0" deprel=""/>
  <word id="7" form="dezvoltarea" lemma="dezvoltare" postag="Ncfsry" head="0" deprel=""/>
  <word id="8" form="literaturii" lemma="literatură" postag="Ncfsoy" head="0" deprel=""/>
  <word id="9" form="şi" lemma="şi" postag="Crssp" head="0" deprel=""/>
  <word id="10" form="a" lemma="al" postag="Tsfs" head="0" deprel=""/>
  <word id="11" form="limbii" lemma="limbă" postag="Ncfsoy" head="0" deprel=""/>
  <word id="12" form="noastre" lemma="meu" postag="Ds1fsop" head="0" deprel=""/>
  <word id="13" form="literare" lemma="literar" postag="Afpfson" head="0" deprel=""/>
  <word id="14" form="a" lemma="avea" postag="Va--3s" head="0" deprel=""/>
  <word id="10" form="determinat" lemma="determina" postag="Vmp--sm" head="0"/>
  <word id="16" form="crearea" lemma="creare" postag="Ncfsry" head="0" deprel=""/>
```

Fişierul rezultat după această procesare a fost transformat în fişier *.xml* printr-un program în PERL, care face modificările necesare astfel încât fişierul să fie adaptat pentru adnotarea sintactică.

Deoarece utilizarea unui adnotator sintactic implică un proces îndelungat şi câteodată anevoios, este necesară conceperea unei metodologii de adnotare la nivel sintactic a corpusului. Această metodologie reprezintă un alt pas în procesul creării.

Avem nevoie de instrucţiuni care să permită o adnotare la nivel sintactic consistentă cu o teorie lingvistică. Până în prezent s-a elaborat o listă cu posibilii regenţi, posibile cuvinte subordonate, relaţii etc. În această listă se mai poate completa pe măsură ce noi exemple sunt descoperite pe parcursul procesului de adnotare sintactică. De exemplu, în figura de mai jos (primul rând) avem un substantiv ca regent. S-a găsit o propoziţie cu un grup nominal în care substantivul era urmat de o prepoziţie şi un adverb. În arborele de dependenţă, pe săgeata dintre substantiv (*plimbarea*) şi prepoziţie (*de*), vom avea notată, aşadar, relaţia de atribut adverbial (*a.adv.*), ca în exemplul: *Plimbarea de azi*. Pentru continuarea analizei acestui grup nominal, trebuie să căutăm în listă situaţia în care regentul este o prepoziţie (*de*) şi cuvântul subordonat este adverb (*azi*) şi să aplicăm aceeaşi metodologie pentru a completa arborele.

REGENT	CUV. SUBORDONAT	URMAT DE	RELATIE	ABREVIERE	EXEMPLU
1. Substantiv	prepozitie	adverb	atribut adverbial	a. adv.	Plimbarea de azi
2. Substantiv	prepozitie	verb nefinit	atribut verbal	a. verb.	Masina de spalat
3. Substantiv	prepozitie	nominal	atribut substantival	a. subst.	Praf de puşcă
4. Substantiv	numeral	-	atribut adjectival	a. adj.	Doi oameni
5. Substantiv	articol	-	determinant	det.	Un om
6. Substantiv	adjectiv	-	atribut adjectival	a. adj.	Fată tânără
7. Substantiv	pronume	-	atribut adjectival	a. adj.	Casa lui
8. Substantiv	verb	-	atribut verbal	a. vb.	Ideea de a merge
9. Adjectiv	prepoziţie	-	complement de agent	c.ag.	Construit de Ion
10. Adjectiv	adverb	-	compl. circum.de mod	c.c.m.	construit repede

Figura 1: Fragment din lista relaţiilor de dependenţă (în curs de dezvoltare)

După operaţiile de preprocesare, se poate trece la cel din urmă pas şi cel mai important pentru proiectul de faţă: adnotarea corpusului la nivel sintactic. Această adnotare manuală, dacă este lipsită de reprezentarea grafică a arborilor de dependenţă, devine un proces foarte dificil. Prin urmare, a fost nevoie de utilizarea unui instrument grafic interactiv cu care să se adnoteze, vizualizeze şi modifice arborii rezultaţi în urma parsării.

TreeAnnotator este un program[6] de editare de arbori de dependenţă funcţională care oferă utilizatorului o interfaţă grafică.

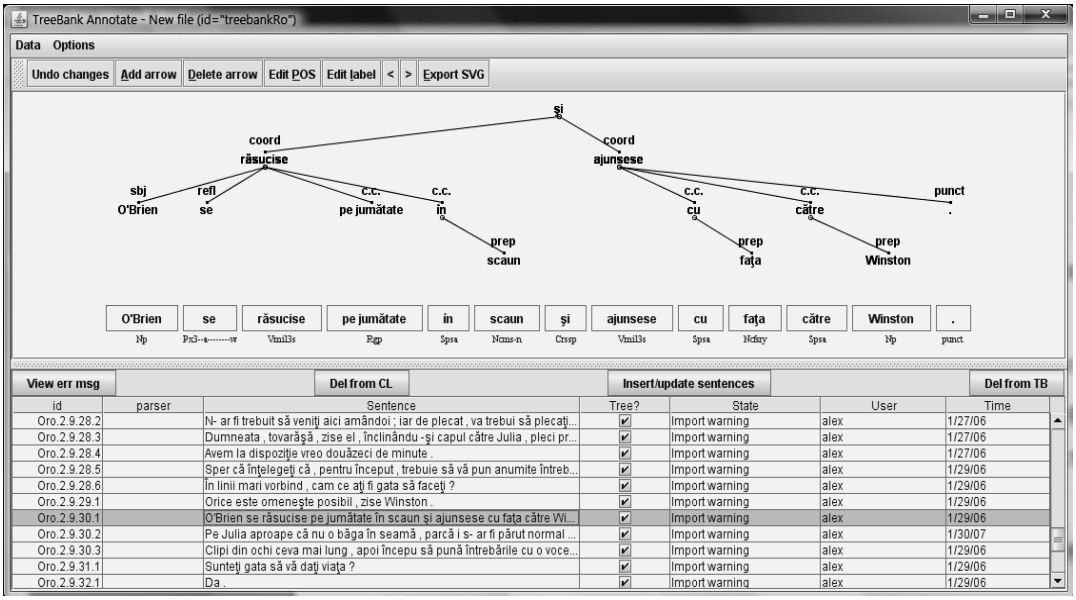


Figura 2: Arbore de dependenţă vizualizat cu TreeAnnotator

Interfaţa permite vizualizarea fiecărui cuvânt al frazei cu eticheta morfologică corespunzătoare. Această facilitate a fost introdusă pentru corectarea eventualelor probleme de adnotare a părţii de vorbire. Pentru realizarea arborelui de dependenţă a

[6] Acest program a fost dezvoltat în cadrul Institutului de Informatică Teoretică Iaşi şi se găseşte la adresa: http://students.info.uaic.ro/~mmoruz/FDAnnotator/TreeAnnotator.tgz.

unei fraze se selectează cuvântul părinte (regentul) urmat de cuvântul fiu (subordonatul), după care se inserează drumul de la părinte către fiu. În urma realizării tuturor legăturilor de dependență dintre cuvinte se obține arborele de dependență.

În continuare vom descrie diferitele tipuri de probleme ivite.

3. Probleme tehnice

S-a întâlnit problema segmentării greșite a unor fraze din cauza punctelor de suspensie sau a parantezelor pătrate cu puncte de suspensie. De exemplu, o frază ca:

„A cerceta evoluția scrierii chilirice românești, începînd din secolul al XVI-lea pînă spre mijlocul secolului al XIX-lea [...], înseamnă, evident, a-i urmări, și cronologic și pe regiuni istorice, mersul și dezvoltarea, precizînd momentele, direcția, semnele și proporția schimbărilor grafice, ca tot atîtea etape dialectice ale acestei forme de cultură, deosebit de importante."[7]

a fost segmentată în două fraze, una conținând textul: *„A cerceta evoluția scrierii chilirice românești, începînd din secolul al XVI-lea pînă spre mijlocul secolului al XIX-lea [...]"* și cealaltă având textul: *„înseamnă, evident, a-i urmări, și cronologic și pe regiuni istorice, mersul și dezvoltarea, precizînd momentele, direcția, semnele și proporția schimbărilor grafice, ca tot atîtea etape dialectice ale acestei forme de cultură, deosebit de importante."* În realitate cele două întinderi constituie o frază doar luate împreună.

Soluția găsită acestei probleme a fost să modificăm textele inițiale și să scoatem punctele de suspensie sau parantezele pătrate cu puncte de suspensie din structura originală a frazei, deoarece acestea nu schimbă cu nimic analiza sintactică a textului, și apoi să trecem frazele prin marcarea automată cu informații lingvistice.

Similar s-au rezolvat și problemele: texte fără punct la sfârșitul propoziției / frazei, ghilimele, propoziții / fraze ce încep cu literă mică.

Deoarece programul adnotator, TreeAnnotator, nu permite adnotarea de noduri izolate, fără legături cu restul nodurilor din frază, iar orice semn de punctuație constituie, la rândul lui, un nod, s-a pus problema relaționării punctului de la sfârșitul frazei / propoziției. De cine trebuie el legat? Am hotărât să legăm punctul de verbul predicativ al ultimei propoziții principale, deoarece un nod nu poate avea mai mulți părinți.

Ne-am mai lovit, totodată, de câteva probleme de procesare (unele texte produceau erori pe parcursul rulării diferitelor programe) și de timp (unele procesări au consumat resurse de timp semnificative), însă în capitolul următor vom descrie unele probleme de natură gramaticală.

4. Probleme gramaticale

Un rol important în analiza gramaticală îl joacă flexiunea, iar prin flexiune ne referim la totalitatea schimbărilor suferite de un cuvânt în aspectul lui formal pentru exprimarea raporturilor sintactice din vorbire între obiectele (ființe și lucruri) din realitatea

[7] Text preluat din N. Anghelescu Temelie, Silviu Constantinescu – *Analize gramaticale*, Ed. Științifică și Enciclopedică, București, 1985, p.10-99;

înconjurătoare, precum şi între obiecte şi acţiuni. Prin urmare, substantivul sau substitutul său (numele obiectului) şi verbul (cuvântul care exprimă acţiunea, starea sau existenţa privite ca proces) îşi vor modifica forma pentru a putea exprima raporturile sintactice care reflectă raporturile existente în realitatea înconjurătoare. Analiza trebuie să se facă deci, în strânsă legătură cu funcţia pe care o îndeplinesc în propoziţie aceste părţi de vorbire. Plecând de la importanţa funcţiei şi de la faptul că un nod nu poate avea mai mulţi părinţi, am întâmpinat problema elementului predicativ suplimentar. Dacă luăm ca exemplu propoziţia „*Azi, ca un sfînt dintr-o icoană veche,*

Blînd îmi răsai cu faţa ta blajină", cuvântul *blînd*, fiind considerat element predicativ suplimentar, ar putea avea ca regent şi verbul *răsai* (cum răsai? blând) şi substantivul *sfînt* (ce fel de sfânt? blând). Astfel că, am hotărât să legăm cuvântul respectiv doar de verb şi să-i atribuim funcţia sintactică de complement circumstanţial de mod.

O altă problemă este cea a conjuncţiilor subordonatoare fără funcţie sintactică. Ele nefiind analizate gramatical în analizele tradiţionale ale unor specialişti, s-a pus problema de cine le legăm şi ce relaţie le atribuim. Deoarece conjuncţia respectivă introduce o propoziţie, ni se pare firesc să fie legată de verbul predicativ al propoziţiei respective, iar relaţia de pe arcul dintre verb şi conjuncţie să fie cea de particulă, conjuncţia fiind considerată, aici, o particulă a verbului respectiv. În ceea ce priveşte conjuncţiile coordonatoare, acestea pot fi noduri părinte (chiar la primul nivel) în arborele sintactic din adnotarea propriu-zisă, fiindcă ele leagă două propoziţii de acelaşi fel, dar nu ţin numai de una sau de cealaltă.

Cea mai des întâlnită problemă a fost cea a locuţiunilor de orice fel. Programele folosite la prepararea textelor (vorbim despre procesele de pre-adnotare), în foarte multe cazuri, nu recunoşteau locuţiunile ca pe un singur nod, ci constituia câte un singur nod pentru fiecare cuvânt inclus în locuţiune. De exemplu, *pas cu pas*, o locuţiune adverbială, este recunoscută de program în trei noduri separate *pas*, *cu* şi *pas*, ceea ce pune în dificultate analiza noastră sintactică, care de altfel ar fi fost foarte simplă dacă am considera locuţiunea ca fiind un singur nod. Pentru a simplifica lucrurile, am umblat, din nou, de data asta, la textul din *.xml* şi am legat cuvintele din locuţiune prin caracterul „underscore" („_") pentru ca programul să recunoască şirul de cuvinte *pas_cu_pas* ca pe un singur nod. Bineînţeles a trebuit să modificăm şi denumirea părţii de vorbire, ea fiind acum, în cazul de faţă, cea de adverb.

5. Concluzii

În concluzii, după ce am trecut în revistă unele probleme şi soluţiile lor, nu ne rămâne decât să tragem linia şi să punem pe masă ceea ce s-a realizat până în prezent. S-a pornit de la o colecţie de aproximativ 2000 de texte (fraze şi propoziţii), însumând în jur de 70.000 de cuvinte şi s-a ajuns la un număr de 15 fişiere *.xml*, care conţin peste 1.800 de fraze adnotate cu aplicaţia TreeAnnotator, având un total de aproximativ 67.000 de cuvinte.

Această colecţie de texte adnotate constituie doar un început, un punct de pornire pentru a demara învăţarea, la rândul nostru, a unui parser sintactic pentru limba română. Vom continua cu adnotările la nivel sintactic, dar de această dată automat cu ajutorul aplicaţiei MALT (Nivre et al., 2007), care creează un model de adnotare la dependenţă pentru limba română, până când se va atinge un număr semnificativ de cuvinte,

deoarece, cu cât programul se va antrena pe un număr cât mai mare de texte, cu atât el va avea o bază de date mai bogată şi rezultatele, în urma unor teste, vor fi mai eficiente.

Un corpus adecvat, de mai multe milioane de cuvinte în uz este dincolo de puterea de organizare a unui singur cercetător, el necesitând o activitate în cadrul unui colectiv. De aceea, corpusul, realizat până în acest stadiu din cercetarea de faţă, constituie un „start" promiţător pentru ceea ce va urma în viitorul apropiat, şi anume un corpus *gold*. Această denumire o poartă doar acele corpusuri ce se supun unor condiţii de realizare, şi anume: corpusurile trebuie să conţină un număr foarte mare de cuvinte (de ordinul milioanelor de cuvinte) care să fie adnotate de cel puţin doi adnotatori umani (bineînţeles, aceştia trebuie să fie specialişti în domeniu).

Referinţe bibliografice

A. Izvoare şi lucrări de referinţă

Anghelescu Temelie, N., Constantinescu, S. (1985). *Analize gramaticale*, Editura Ştiinţifică şi Enciclopedică, Bucureşti, p.10-99.

Anghelescu Temelie, N., Popescu, A. G. (1976). *Dificultăţi ale analizei gramaticale*, Editura Ştiinţifică şi Enciclopedică, Bucureşti, p.131-291.

Botiş, V., Vulişici Alexandrescu, M., Comănescu, I. (1977). *Sintaxa propoziţiei*, Editura Facla, p. 50-103.

Constantinescu, S. (1976). *Exerciţii şi analize gramaticale*, Editura Didactică şi Pedagogică, Bucureşti, p.61-80.

Drăşoveanu, D. D. (1959-1966). *Analize gramaticale şi stilistice*, Editura Ştiinţifică, Bucureşti, p. 111-155.

Lupu, C., Vlădescu, A. (1997). *Limba română prin exerciţii*, Editura Logos, Bucureşti, p. 18.

Metea, A., Drincu, S. (1974). *Analize gramaticale*, Editura Facla, Timişoara, p.23-109.

Muţiu, I., Bercea, L. P. (1985). *Exerciţii de gramatică*, Editura Facla, Timişoara, p.33-77.

Vlad, V., Ştirbu, P., Vlad-Budoiu, V. (1978). *Analize gramaticale structurale*, Editura Dacia, Cluj-Napoca, p.57-550.

B. Literatură secundară

Academia Română (2005). *Gramatica limbii române*, Editura Academiei Române, Bucureşti.

Călăcean, M. and Nivre, J. (2009). A Data-Driven Dependency Parser for Romanian. In Proceedings of TLT-7.

Curteanu, N., Gâlea, D., Butnariu, C. (2003). *Segmentation Algorithms for Clause-Type Textual Units*, versiunea în limba română în „Limba română în societatea informaţională", D. Tufiş, Fl. Filip Eds., Romanian Academy, Bucharest, Expert Press, p.165-190.

Curteanu, N., Moruz, A., Trandabăţ, D., Bolea, C., Dornescu, I. (2006). The Structure and Parsing of Romanian Verbal Group and Predicate, Advances in Intelligent

Systems and Technologies ECIT2006 – 4th European Conference on Intelligent Systems and Technologies, Iasi, Romania, Septembrie 21-23, pp. 93-105.

Hudson, R., (1998). English Grammar. London, Rontledge.

Mel'Cuk, I. A., (1987). Dependency Syntax: Theory and Practice. Buffalo, Suny Press.

Nivre, J. (2007). Data-Driven Dependency Parsing across Languages and Domains: Perspectives from the CoNLL 2007 Shared Task. In Proceedings of the Tenth International Conference on Parsing Technologies, 168-170.

Seretan, V., Wehrli E., Nerima L., Soare G. (2010). (submitted for publication) FipsRomanian : Towards a Romanian Version of the Fips Syntactic Parser. LREC 2010.

Stati, S. (1972). *Elemente de analiză sintactică*, Editura Didactică şi Pedagogică, Bucureşti.

Tufiş, D., Barbu, A. M., Pătraşcu, V., Rotariu, G., Popescu, C. (1997). *Corpora and Corpus-Based Morpho-Lexical Processing*, in „Recent Advances in Romanian Language Technology", Dan Tufiş, Poul Andersen, Editura Academiei Române.

Tufiş, D. (2000). *Using a large set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging*, Proceedings of LREC.

http://www.di.unito.it/~tutreeb/

http://consilr.info.uaic.ro/ro/index.php?showpage=030102

http://www.ceid.upatras.gr/Balkanet/publications.htm

http://www.sketchengine.co.uk/

http://ufal.mff.cuni.cz/pdt2.0/

http://www.acad.ro/pro_pri/doc/st_h01.doc

http://students.info.uaic.ro/~mmoruz/FDAnnotator/TreeAnnotator.tgz

http://www.racai.ro/webservices/TextProcessing.aspx

MONITORIZAREA PRESEI ÎN CADRUL PROIECTULUI NEOROM

ANA-MARIA BARBU

Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti", Bucureşti – România

anamaria.barbu@g.unibuc.ro

Rezumat

Acest articol prezintă contribuţia părţii române în cadrul proiectului Neorom. Proiectul are ca obiectiv dezvoltarea unei baze de date cu neologismele lexicale adunate din ziare, începând cu anul 2004, în toate limbile romanice. Articolul are trei părţi. În prima parte este prezentat sistemul românesc de monitorizare a presei, în a doua, interfaţa Neorom de înregistrare a neologismelor în baza de date, iar în a treia, o scurtă analiză a cuvintelor din bază din perspectiva modului lor de formare.

1. Introducere

Proiectul Neorom (proiect de Neologie romanică) este un proiect internaţional care are ca obiectiv crearea unei baze de date multilingve în care sunt înregistrate cuvintele nou apărute în presa scrisă, în paralel pentru următoarele limbi romanice: catalană, franceză (din Franţa, Belgia şi Québec), galiciană, italiană, portugheză (din Portugalia şi Brazilia), română şi spaniolă. Inventarierea cuvintelor a început în anul 2004 şi continuă în fiecare an.

Acest proiect se desfăşoară sub egida organizaţiei Realiter (http://www.realiter.net) şi a Uniunii Latine (http://www.unilat.org) şi este condus de Maria Teresa Cabré, profesor la Institutul pentru Lingvistică Aplicată din cadrul Universităţii Pompeu Fabra, Spania. El este concretizat printr-o platformă de căutare a cuvintelor inventariate în bazele de date corespunzătoare limbilor implicate (Cabré 2004). Platforma Neorom se găseşte la adresa http://obneo.iula.upf.edu/bneorom/index.php.

Proiectul Neorom este unul dintre proiectele grupului L'Observatori de Neologia (http://www.iula.upf.edu/obneo/obpresca.htm) cu cea mai importantă activitate în domeniul neologiei şi al terminologiei. Mai putem menţiona ca proiecte similare proiectul Neoscope (http://www.certa.usj.edu.lb/files/neoscope.htm) şi secţiunea de achiziţie de cuvinte noi *Oxford English Dictionary*, http://dictionary.oed.com/about/.

În accepţia proiectului Neorom, cuvintele nou apărute, numite şi *neologisme*[1], sunt acele unităţi lexicale formate din unul sau mai multe cuvinte (grafice) care nu sunt înregistrate într-un corpus prestabilit de dicţionare reprezentative ale limbii considerate, numit *corpus de excludere*. Prin urmare, cuvintele inventariate în baza românească Neorom nu sunt neapărat cele care apar în presă strict într-un anumit an, ci şi acelea care dintr-un motiv sau altul nu au fost înregistrate în dicţionare, deşi sunt folosite în limbă de mai

[1] Termenul de *neologism* nu are o semnificaţie unanim acceptată, de aceea pentru unii lingvişti folosirea lui în contextul acestui proiect poate părea improprie. O dezbatere asupra semnificaţiei acestui termen ar depăşi spaţiul acestui articol.

multă vreme. Pe de altă parte, proiectul acordă deopotrivă atenţie cuvintelor care nu au intrat propriu-zis în limbă şi e posibil să nu intre niciodată, dar reprezintă mărturii asupra mijloacelor de îmbogăţire a limbii şi asupra inventivităţii lingvistice. Altfel spus, nu se aplică o selecţie a cuvintelor în funcţie de numărul lor de apariţii în corpusul de ziare considerat, ci se inventariază până şi hapaxurile. De asemenea, trebuie precizat că în baza de date Neorom înregistrăm şi cuvinte care par să existe în dicţionare, dar care au suferit totuşi o modificare; de pildă, şi-au schimbat partea de vorbire sau genul, au căpătat forme de plural sau folosesc alte morfeme de plural, se scriu altfel (cu cratimă sau fără cratimă sau ca în limba de origine sau reprezintă greşeli de ortografie) ş.a.

Cercetarea prezentată în acest articol constă în dezvoltarea bazei de date româneşti din cadrul proiectului Neorom. În secţiunea următoare descriem sistemul de monitorizare folosit pentru inventarierea cuvintelor ce urmează a fi introduse în bază. În secţiunea 3 prezentăm interfaţa bazei de date Neorom cu câmpurile din fişa care trebuie completată pentru fiecare cuvânt introdus. Unele dintre câmpuri constituie criterii de căutare în bază, cunoaşterea lor fiind necesară celor care vor dori să utilizeze baza Neorom. Secţiunea 4 o dedicăm unei scurte analize a rezultatelor româneşti din cadrul proiectului. Utilitatea acestui proiect şi în special a bazei de date româneşti este subliniată în capitolul de concluzii.

2. Sistemul de monitorizare

a. Date preliminare

Corpusul folosit pentru monitorizarea cuvintelor nou-apărute în presă este alcătuit din mii de articole (însumând circa 14 milioane de cuvinte) culese de pe Internet din variantele on-line a mai multor ziare precum *Adevărulonline, BBC-România*, *Cotidianul, Jurnalul Naţional, Libertatea, România Liberă* şi altele, apărute în perioada 01.08.2004 – 15.10.2007. Corpusul este accesibil prin bunăvoinţa unor colegi italieni. Trebuie subliniat faptul că obţinerea textelor dintr-un ziar on-line, într-un format care poate fi prelucrat automat ulterior, nu este deloc un lucru trivial, de aceea ne-am bucurat să putem beneficia de ajutorul colegilor noştri. Începând cu anul acesta vom încerca să folosim programul Buscaneo, pus la dispoziţie de colegii spanioli ai acestui proiect, însă avem reticenţe în privinţa eficienţei lui. Pentru fiecare cuvânt din ziarul inspectat pe care programul nu-l găseşte în dicţionarul de forme flexionare, Buscaneo deschide o fişă Neorom. Dicţionarul nostru electronic fiind relativ mic, numărul de cuvinte negăsite este foarte mare, prin urmare şi numărul de fişe deschise va fi covârşitor şi inutil de mare, deoarece vom fi privaţi de eficienţa lucrului cu liste de cuvinte pe care le putem ulterior prelucra automat sau le putem inspecta vizual.

Este de menţionat că *BBC-România* foloseşte diacritice, celelalte ziare nu folosesc diacritice. Toate ziarele folosesc grafia cu „â” (care în textul fără diacritice apare „a”), cu excepţia ziarului *Cotidianul* care foloseşte şi grafia cu „î” (care în textul fără diacritice apare „i”).

Dicţionarele româneşti care alcătuiesc corpusul de excludere, adică referinţa faţă de care un cuvânt este considerat nou apărut (anume, dacă nu se găseşte deja înregistrat în corpusul de excludere respectiv) sunt cele în format electronic disponibile la adresa

http://dexonline.ro, din care avem în vedere DEX '98, DLRA, DLRC, DLRM, DN, MDA şi MDN, precum şi DOOM2, pentru că prezintă avantajul căutării semiautomate.

b. Procesul de monitorizare

Modul în care selectăm cuvintele nou apărute este unul relativ simplu, după cum se poate vedea în figura 1, însă destul de laborios prin volumul mare de date şi prin combinarea etapelor automate, semiautomate şi manuale. Articolele din ziare sunt indexate cu ajutorul concordanţierului LUCON (http://sourceforge.net/projects/lucon/). Cuvintele (flexionate) din indexul obţinut sunt căutate automat în dicţionarul electronic de forme flexionare RoMorphoDict (Barbu, 2008), iar cele care nu au fost găsite în acest dicţionar (care conţine paradigmele complete ale celor aprox. 60000 de cuvinte-titlu înregistrate în DOOM 1) sunt trecute într-o listă de cuvinte „necunoscute". Această listă poate fi impresionant de mare, de ordinul zecilor de mii de cuvinte. O formă necunoscută poate fi înregistrată astfel din următoarele motive:

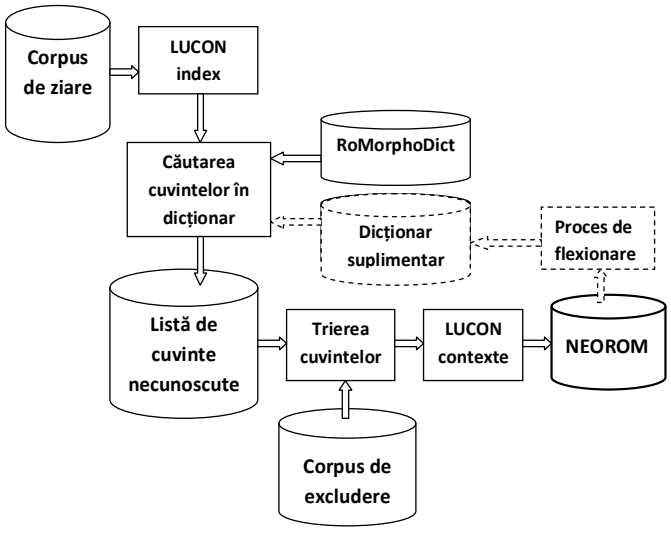


Figura 1: Sistemul de monitorizare

- forma respectivă nu se află în dicţionarul electronic folosit ca referinţă (pentru că acesta conţine doar 60000 de leme, multe dintre acestea nefiind folosite în limbajul curent);
- face parte dintre abrevieri, unităţi de măsură sau alte simboluri;
- este un nume propriu;
- este un cuvânt străin;
- reprezintă o greşeală tipografică sau a vorbitorilor (de pildă, forma *ostatec* este întâlnită de 111 ori);
- exprimă perioade şi aproximări (de exemplu, ianuarie-aprilie, primăvară-vară, opt-nouă).

Această listă de cuvinte „necunoscute" este apoi inspectată vizual în etapa de triere a cuvintelor, pentru a exclude acele cuvinte pe care le socotim reziduale, iar ceea ce rămâne este confruntat, de asemenea, manual, cu dicţionarele amintite din corpusul de

excludere, care au totuşi avantajul că sunt introduse într-o bază de date accesibilă pe Internet. Pentru fiecare cuvânt care nu a fost găsit în corpusul de excludere se completează o fişă de introducere în baza de date Neorom, după cum e prezentat în secţiunea următoare.

Procesul descris până aici trebuie completat cu o nouă etapă în care cuvintele care au fost introduse în baza Neorom sunt flexionate şi înregistrate într-un dicţionar suplimentar de referinţă, pentru a nu mai fi inventariate încă o dată. De fapt întregul proces poate fi îmbunătăţit dacă am avea la dispoziţie un program de descărcare a textelor de ziare de pe Internet, dacă corpusul de excludere ar fi în formă electronică corespunzătoare sau dacă am dispune de un program de flexionare a cuvintelor necunoscute pe care să-l putem integra într-un proces de prelucrare.

3. Interfaţa platformei Neorom

În această secţiune descriem conţinutul fişei care trebuie completată pentru fiecare cuvânt nou introdus în bază. După cum se vede în figura 1, câmpurile care trebuie completate sunt următoarele.

Date generale despre **despuiere**:

Tipo – Se indică faptul că datele sunt extrase din presa scrisă.

Lengua – Datele culese sunt din limba română.

Ámbito geográfico – Este vorba de limba română vorbită în România.

Fuente – Ziarul din care s-a extras cuvântul respectiv este *Libertatea* (varianta electronică).

F. Publicación – Data ediţiei ziarului este 06 ianuarie 2005.

Datele fişei:

Entrada – Se indică cuvântul-titlu.

Categoría gramatical – Categoria gramaticală (partea de vorbire, genul, numărul, tipul verbului etc.) se alege dintr-o listă predefinită. În privinţa categoriei gramaticale, ne-am confruntat la început cu faptul că în lista predefinită nu exista genul neutru (deoarece nu există în celelalte limbi romanice), însă ulterior această problemă a fost rezolvată.

Contexto – Pentru fiecare cuvânt-titlu se dă contextul în care apare în numărul respectiv al ziarului. În cele mai multe cazuri, un cuvânt apare în mai multe numere ale mai multor ziare. În această situaţie se alege de obicei numărul cel mai vechi, însă acolo unde este cazul se preferă contextul care sugerează cel mai bine sensul cuvântului respectiv (de pildă, sunt contexte în care se dă chiar o scurtă definiţie a cuvântului respectiv). Contextele sunt obţinute cu ajutorul concordanţerului Lucon, care, pe lângă multe alte facilităţi, o are pe aceea că oferă o cale directă şi simplă de obţinere a sursei contextului, adică ziarul şi ziua ediţiei (date ce trebuie completate în fişă).

Aspectos tipográficos – Această rubrică se referă la modalităţile de subliniere a unui cuvânt. În cele mai multe cazuri, sublinierea se face cu ghilimele sau paranteze, însă în ediţiile pe hârtie pot exista şi alte modalităţi.

Tipo de neologismo – Acest câmp este destinat indicării modului de formare a cuvântului nou. Se alege dintr-o listă predefinită simbolul corespunzător modului de formare a cuvântului. Conţinutul listei predefinite, precum şi o analiză detaliată a rezultatelor sunt prezentate în (Barbu 2010).

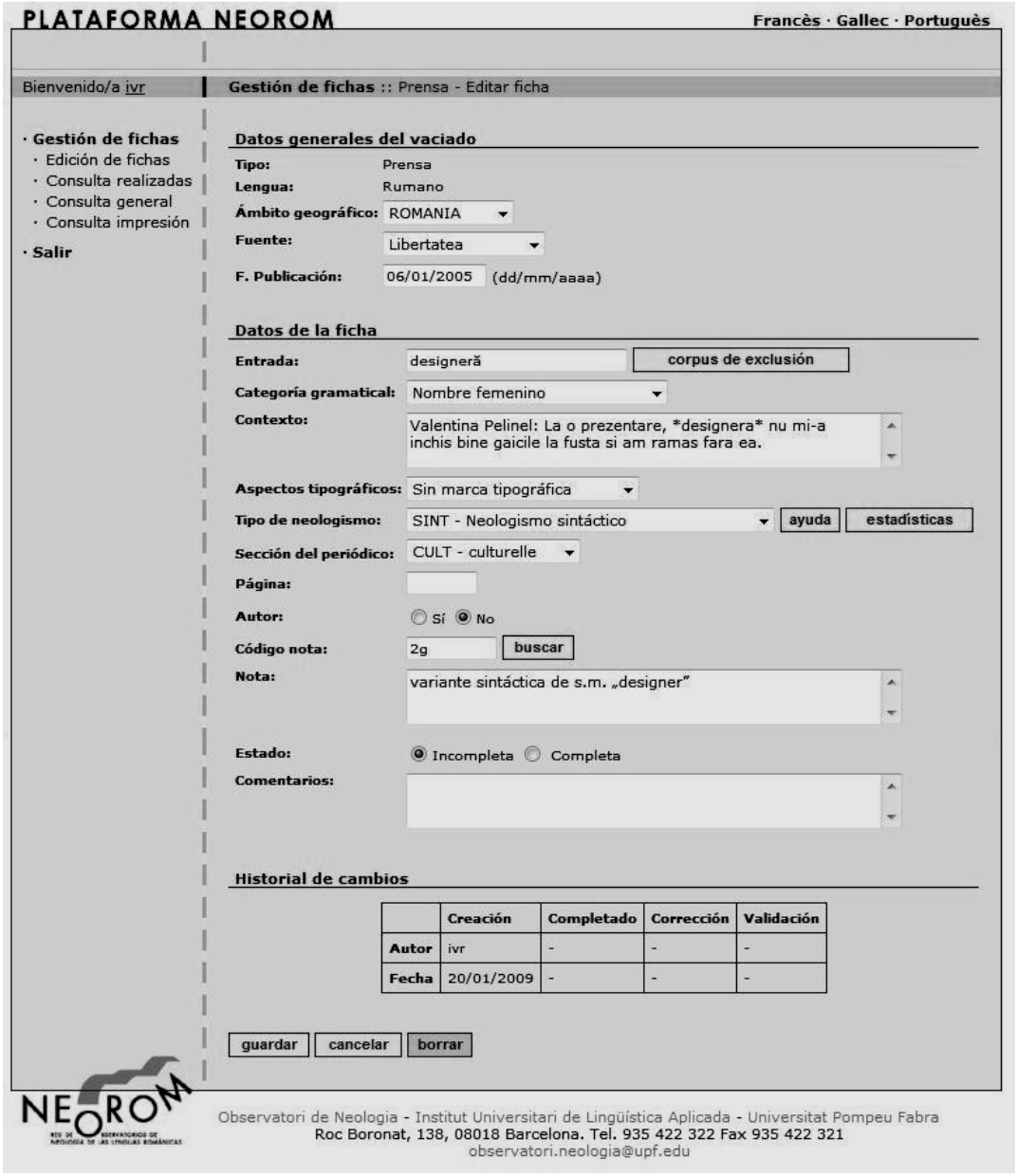


Figura 2: Fişa unei intrări din baza de date Neorom

Sección del periódico – Se indică secţiunea articolului din care a fost extras cuvântul. Domeniile cărora le aparţin în general secţiunile unui ziar sunt cel politic, economic, financiar, social, comercial de autoturisme, economic, informatic-electronic, sănătate, sport, cultural şi de diverse. Pentru că în corpusul electronic folosit articolele nu sunt organizate pe secţiuni, domeniul este dedus după conţinutul articolului.

Página – În cazul ediţiilor pe hârtie se poate indica pagina publicaţiei, însă noi, folosind ediţiile elctronice, nu avem acces la această informaţie.

Autor – Dacă se cunoaşte autorul articolului sau cine a utilizat cuvântul (în cazul unui citat), acest fapt poate fi indicat.

Código nota – Pentru anumite tipuri de neologisme se poate adăuga un cod de notă explicativă, ales dintr-o listă predefinită.

Nota – Acest câmp conține nota explicativă propriu-zisă. În exemplul dat în figura 1, cuvântul *designeră* este un neologism sintactic (pentru definiție vezi secțiunea următoare), pentru că este obținut prin schimbarea genului de la substantivul masculin *designer* la unul feminin.

Estado – Starea fișei poate fi precizată în această rubrică prin indicarea uneia dintre valorile Incompletă/Completă.

Comentarios – Colaboratorii, care sunt membri ai proiectului, pot face diferite comentarii pe marginea fișei respective, ele nefiind făcute publice.

Historial de cambios – Zona de evidență a modificărilor operate în fișă este completată automat de sistem cu numele colaboratorului care a creat, a completat, a corectat sau a validat fișa și cu data operației respective.

Procesul de realizare a fișei se încheie apăsând butonul de salvare 'Guardar' sau lăsând fișa neschimbată prin butonul de anulare 'Cancelar'. Ștergerea fișei din baza de date se face prin butonul 'Borrar'.

4. Scurtă analiză a rezultatelor

Până în momentul de față au fost introduse în baza Neorom circa 2000 de cuvinte, urmând ca până la sfârșitul anului 2010 inventarul să fie extins la peste 4000. În faza actuală, completarea bazei de date se poate face într-un ritm de 100 de intrări pe zi. Clasificarea modului de formare a cuvintelor a trebuit să-l facem după standardul propus în cadrul proiectului, comun pentru toate limbile implicate. Pentru a ne alinia la acest standard, a fost necesar să facem unele compromisuri prezentate în cele ce urmează.

- Clasificarea cuvintelor după modul de formare se bazează pe forma lor analizabilă, iar nu pe etimologia lor în mod strict (de pildă, unele dintre cuvintele analizabile pot fi de fapt împrumuturi).
- Dintre mai multe moduri de formare posibile s-a ales unul singur. În principiu, s-a ținut seama de cuvintele deja înregistrate în dicționare ca bază de derivare. De exemplu, cuvântul nou *coprezentatoare* a fost considerat ca reprezentând derivarea cu prefix a lui *prezentatoare*, mai degrabă decât sufixarea moțională de la *coprezentator*, care nu e nici el înregistrat. Alt exemplu: un cuvânt precum *sepepist* e format atât prin siglare, cât și prin sufixare, dar s-a înregistrat în baza Neorom doar format prin siglare.
- Anumite moduri de formare pot diferi de ceea ce propune literatura de specialitate românească. De pildă, în proiectul Neorom femininul *designeră* este considerat un neologism sintactic (sau gramatical), în timp ce în literatura românească el este un derivat prin sufixare. Un alt exemplu este acesta: dacă înlăturând fie prefixul, fie sufixul aceluiași cuvânt obținem tot câte un neologism atunci cuvântul este considerat un derivat prin 'sufixare și prefixare' în Neorom (ex. *superblatist*: atât *superblat*, cât și *blatist* sunt neologice), pe când în literatura românească asemenea derivate nu sunt decât cele în care sufixul și prefixul sunt atașate unei baze inexistente ca atare în limbă (ex. *îm+brăc+a*).

- Proiectul Neorom consideră că prefixoide care în limba de origine au fost prepoziții sau adverbe (ex. *anti-, extra-, intra-*) formează derivate prin prefixare, iar celelalte prefixoide sunt elemente de compunere cultă (ex. *auto-, agro-, bio-*).
- Din păcate, proiectul Neorom nu oferă posibilitatea notării cuvintelor formate prin substituție de sufixe (ex. *atotputere* < *atotputernic*) sau cele prin derivare regresivă (*bâlbă* < *bâlbâi*).

Pentru stabilirea modurilor de formare, pe lângă recomandările din cadrul proiectului, am folosit și lucrările (Coteanu et al., 1985) și (Stoichițoiu-Ichim, 2001). Precizăm însă că modul de formare indicat pentru fiecare cuvânt din bază nu constituie rezultatul unei cercetări riguroase de etimologie, fapt care lasă loc de studiu specialiștilor interesați.

În cele ce urmează este prezentată succint repartizarea cuvintelor din baza de date românească din acest moment, în funcție de modurile de formare cele mai reprezentative (pentru care indicăm simbolurile folosite în platforma Neorom).

Derivare cu afixe: 27,5%

Derivare cu sufix – FSUF: 13,54%. Cele mai frecvente sufixe sunt **-ist** (ex. *cablist* „lucrător la cablu", *galerist* „proprietar de galerie de artă", *forumist* „utilizator de forumuri"), **-re** (ex. *fumare, gudurare, microcipare*) și **-(e/i)an** (ex. *becalian*, *cațavencian*). Cel mai neobișnuit sufix este **-ing** (ex. *bancheting* „organizare de banchete", *cuponing* „instrument de marketing ce presupune folosirea cupoanelor").

Derivare cu prefix – FPRE: 13,96%. De departe cel mai des utilizat prefix este **ne-** (30% din derivatele cu prefix) (ex. *nealegere, necalitativ, nesustenabil*) urmat îndeaproape de prefixoidul prepozițional **anti-** (22,5%), care este frecvent folosit pentru a forma adjective din substantive (ex. *anticutremur, antisărăcie, antisforăit*). Un alt prefix îndrăgit de vorbitori este **cvasi-** (ex. *cvasi-absență*, *cvasinațional*, *cvasi-război*).

Compunere: 27,9%

Compunerea din cuvinte întregi – FCOM: 17%. Acest mod este folosit cu precădere în exprimarea relațiilor politice, economice sau sociale (ex. *americano-canadian, chino-ruso-indian, etnico-religios*), a relațiilor economice sau organizatorice (ex. *administrativ-teritorial, cerere-ofertă, financiar-contabil*), în exprimarea numelor de funcții sau meserii (ex. *director-coordonator, cazangiu-tinichigiu*) și altele.

Compunerea cultă (sau savantă) – FCULT: 8,8%. Acest tip de compunere se face cu ajutorul prefixoidelor (altele decât foste prepoziții sau adverbe) dintre care, în mod categoric, cel mai bine reprezentat este prefixoidul neologic **euro-** (ex. *euroasistat, eurobarometru, eurogară, europubelă*). Dintre prefixoidele clasice este folosit cu precădere **auto-,** cu sensul „de la sine, pe sine" (ex. *autoadresa, autobronzant, autoironic*). De remarcat este, de asemenea, apariția elementelor noi, de origine engleză, ***tech-*** (ex. *tech-cultură, tech demo*) și ***web-*** (*web designer/web-designer, web hosting, website*).

Compunerea prin abreviere (siglare – FTSIG, acronimie – FTACR, trunchiere – FTABR): 2,1%. Acest fel de compunere înregistrează două manifestări principale. Una este aceea de combinare a siglării numelor de partide sau organizații cu sufixul **-ist** pentru desemnarea membrilor grupării respective: *cederist, cekist, sepepist/sepepistă, udemerist*. Foarte rar se găsesc cuvinte formate prin siglare pură ca *sereleu* sau *sms*. A

doua manifestare privește trunchierea sau decuparea prefixoidelor din compusele culte pentru a fi folosite ca adjective invariabile sau, apoi, ca substantive. Exemple ale aplicării acestui procedeu înregistrate în baza de date sunt *combo, eco, ego, electro, etno, hidro, homo, lesbi, macro, mix, termo, turbo, vice.*

Conversiune – FCONV: 4,4%

Conversiunea este procedeul prin care un cuvânt este format prin schimbarea părții de vorbire, fără modificarea bazei lexicale. Substantivizarea adjectivelor este, de departe, tipul de conversiune cel mai des întâlnit. Se pot constata câteva direcții principale de manifestare a acestui tip de conversiune.

- când o persoană (sau, mai general, o ființă) este desemnată printr-o calitate a ei, mai ales la plural: *anchetat, mascat, național, necăjit, penal, vestic* etc.
- când persoanele sunt desemnate prin culorile care le definesc ca grup, fie că e vorba de o culoare politică: *verzii*, *alb-albaștri,* fie că e vorba de culorile unei echipe: *alb-violet*, *roșu-albaștri, tricolori, vișinii* etc.
- când se urmărește denumirea unei trăsături prin ceea ce are generic, caz în care adjectivul care denumește trăsătura respectivă este convertit în substantiv la singular, articulat hotărât: *derizoriul, divinul, economicul, eroicul, eroticul, esențialul, eticul*.
- când este suprimat elementul determinat de adjectiv dintr-o sintagmă uzuală: *(alegeri) anticipate, generale, legislative, locale; (aparate) electronice, (medicamente) compensate; (scor) egal* etc.

Neologisme sintactice (sau gramaticale) – SINT: **2.3%**

Neologismele sintactice sunt cele care implică o schimbare de subcategorie gramaticală (gen, număr, regim verbal etc.) într-o bază lexicală. Acestea vizează în special substantivele masculine care capătă forme pentru genul feminin (ex. *administratoră, arbitră, doctoră, fană, stelistă*), substantivele neutre care devin masculine prin forma de plural (ex. *amortizori, boscheți, flotori*) sau substantivele care primesc forme de plural (ex. *bejuri, designuri, disconforturi, euroi*).

Variații ortografice sau flexionare – FVAR: 9,3%

Sunt socotite variații ortografice în special cuvintele care sunt scrise altfel decât în forma normată sau înregistrată în corpusul de excludere: *aiatolah* (normat, *ayatolah), afrodisiac (afrodiziac), chic (șic), click (clic), boutique (butic)*. De asemenea s-au înregistrat numeroase cuvinte în care prefixul sau prefixoidul este despărțit prin cratimă de bază când norma recomandă scrierea legată.

Variațiile flexionare sunt schimbări ale morfemelor flexionare față de cele normate: *aeropoarte* (normat *aeroporturi), cenușei (cenușii), gazoane (gazonuri), girofare (girofaruri).*

Împrumuturi din engleză – ME: 19%

Împrumuturile din engleză ocupă, după cum se vede, aproape o cincime din totalul cuvintelor noi. Se pot distinge câteva domenii în care afluența de anglicisme este remarcabilă:

- nume de profesii: *advertiser, account manager, chief executive officer, chief head-hunter manager, sound-designer, web designer* ș.a.

- sport: *cross country, down hill,four-cross, full-contact, greencard, goal-keeper* ş.a.
- muzică: *backing vocal, boy-band, cover, dance* (adj.)*, (deep) house, death-metal, downtempo, drum & bass, electric-gipsy, featuring, funk, heavy rotation* ş.a.
- modă: *baby doll, face-painting, fashion, fusion-fashion, gloss, hair-styling* ş.a.
- electronică/informatică: *dial-up, downloada, enter, gateway, hands-free, hi-tech* ş.a.
- alimentaţie: *fresh, junk-food, light* ş.a.

Există, firește, şi împrumuturi din alte limbi, însă numărul lor este nesemnificativ. De asemenea sunt înregistrate şi împrumuturi adaptate (adică împrumuturi care nu mai păstrează grafia din limba de origine), dar şi ele sunt puţin numeroase. Numărul lor este cu atât mai mic cu cât s-a constatat o preferinţă a unor vorbitori de a folosi grafia originală (probabil ca pe un act de cultură) în ciuda faptului că cea adaptată a fost deja normată (a se vedea asemenea exemple la rubrica Variaţii ortografice). Exemple pot fi date din franceză: *biju*, (albastru, bleu) *ciel*, *connaisseur* ; chineză: *feng-shui, chi, thai-chi*; italiană: *dolce far niente, famiglia, silenzio stampa, Squadra Azzurra* ş.a.

Xenisme – Xenism: 0,5%

Deşi semnificaţia generală a termenului *xenism* este aceea de împrumut neadaptat dintr-o limbă străină, noi deosebim xenismele de împrumuturile propriu-zise prin aceea că desemnează realităţi care nu sunt specifice zonei în care se vorbeşte limba-gazdă. Xenisme se găsesc, de pildă, în descrierile realităţilor din alte ţări. Asemenea exemple sunt *datsan* („curent budist”), *fatwa* („edict religios”), *kumara* („cultură agricolă din insula Paştelui”), *guerilleros* („luptători de gherilă”)*, mzungus* („om din rasa albă”).

Altele – A: 5%

Pentru unele cuvinte nu se poate stabili etimologia, cum sunt cuvintele simple dialectale, argotice, cazuri ciudate, dificil de etichetat, dar care totuşi sunt cuvinte nou folosite în presă. În această categorie am introdus termenii de specialitate din domeniul farmaceutic, medical etc. (*aflatoxină, botox, carmol*) sau cuvinte care nu au o etimologie evidentă (*amărăştean, amoform, aszu, critifiction, flex* ş.a.).

5. Concluzii

Studiul de faţă atinge mai multe aspecte. În primul rând prezintă prima încercare de monitorizare sistematică a presei scrise româneşti graţie resurselor electronice şi informatice disponibile. Această cercetare poate constitui şi o piatră de hotar pentru o monitorizare ulterioară precisă. Ne referim la faptul că obiectivul proiectului Neorom nu este atât cel de inventariere a neologismelor din presă, cât acela de inventariere a cuvintelor prezente în limbajul presei într-un anumit an precis. Or acest lucru nu se poate realiza, cât de cât exact, fără să existe un moment în care se trage linie şi se spune „iată inventarul cuvintelor existente în limbă la acest moment”. În al doilea rând, acest studiu este important pentru lexicografia românească. S-a constatat că un număr destul de mare de cuvinte nu au fost încă înregistrate în dicţionare în ciuda frecvenţei mari de folosire şi a vechimii lor. Multe dintre acestea reprezintă membri ai unor familii de cuvinte înregistrate parţial (de pildă apare în dicţionar *minimalist*, dar nu şi *minimalism*). De acum această deficienţă poate fi corectată. Există şi situaţii în care cuvintele din bază se găsesc în dicţionare specializate pe domenii (argou, economic, informatică, sport

etc.) sau lingvistice (de sinonime, de antonime sau ortografice), însă faptul că aceste cuvinte apar în presă arată că ele sunt accesibile publicului larg şi de aceea multe dintre ele ar trebui să se regăsească în dicţionarele explicative, dacă depăşesc un anumit prag de apariţii. Trebuie spus că există şi cuvinte care în mod surprinzător nu au fost înregistrate (precum *denim* sau *sclipici*). Chiar dacă nu orice cuvânt din baza de date Neorom se cuvine a fi introdus într-un dicţionar al limbii române, studiul acesta este important şi pentru lexicologie prin reflectarea creativităţii lingvistice, a tendinţelor de evoluţie a limbii precum şi a raportului dintre limba română şi celelalte limbi romanice implicate în proiect.

Mulţumiri. Doresc să mulţumesc d-nei Ioana Vintilă-Rădulescu, coordonatoarea părţii române a proiectului Neorom, pentru sprijinul nemijlocit acordat în îmbunătăţirea calităţii rezultatelor noastre. Mulţumesc de asemenea referenţilor acestei lucrări.

Referinţe bibliografice

Barbu, A.M. (2008). Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. *Proceedings of Language Resources and Evaluation Conference - LREC 2008*, Marrakech şi on-line http://www.lrec-conf.org/proceedings/lrec2008/.

Barbu, A.M. (2010). Baza de date românească din cadrul platformei Neorom. *Limba română*, nr.2, Bucureşti, sub tipar.

Cabré, M. T. (coord.) (2004). *Metodología del trabajo en neología: criterios, materiales y procesos*. Observatori de Neologia, Papers de l'IULA. Sèrie Monografies, 9, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona.

Coteanu, I., Forăscu, N., Bidu-Vrănceanu, A. (1985). *Limba română contemporană. Vocabularul*. Editura Didactică şi Pedagogică, Bucureşti.

Stoichiţoiu-Ichim, A. (2001). *Vocabularul limbii române actuale. Dinamică, influenţe, creativitate*. Editura All, Bucureşti.

Sigle

DEX '98= *Dicţionarul explicativ al limbii române,* Editura Univers Enciclopedic, Bucureşti, 1998.

DLRA = Zorela Creţa, Lucreţia Mareş, Zizi Ştefănescu-Goangă, Flora Şuteu, Valeriu Şuteu, *Dicţionar al limbii române actuale* (ediţia a II-a revăzută şi adăugită), Editura Curtea Veche, Bucureşti, 1998.

DLRC = Vasile Breban, *Dicţionarul limbii române contemporane,* Editura Ştiinţifică şi Enciclopedică, Bucureşti, 1980.

DLRM= *Dicţionarul limbii române moderne,* Editura Academiei, Bucureşti, 1958.

DN = Florin Marcu şi Constant Maneca, *Dicţionar de neologisme,* Editura Academiei, Bucureşti, 1986.

DOOM 1 = *Dicţionar ortografic, ortoepic şi morfologic al limbii române,* Editura Academiei Republicii Socialiste România, Bucureşti, 1989.

DOOM 2 = *Dicţionar ortografic, ortoepic şi morfologic al limbii române,* ediţia a II-a, Editura Univers Enciclopedic, Bucureşti, 2005.

MDA = *Micul dicţionar academic,* Editura Univers Enciclopedic, Bucureşti, 2002.

MDN = Florin Marcu, *Marele dicţionar de neologisme,* Editura Saeculum, Bucureşti, 2000.

EMOȚII ÎN CUVINTE: ELABORAREA RESURSEI MULTILINGVE

VICTORIA BOBICEV, VICTORIA MAXIM, TATIANA PRODAN,
NATALIA BURCIU, VICTORIA ANGHELUȘ

Universitatea Tehnică a Moldovei, Chișinău, Moldova

*vika@rol.md, maxivica@yahoo.com, tatiana.ursulenco@gmail.com,
natusicb@yahoo.com, lazu_vic@yahoo.com*

Rezumat

În lucrarea dată este descris procesul de creare a WordNet-Affect pentru limbile română și rusă. WordNet-Affect reprezintă o resursă lexicală elaborată în baza WordNet-ului englez ce conține informații despre emoțiile pe care le transmit cuvintele. WordNet-Affect este organizat în șase emoții de bază: *anger, disgust, fear, joy, sadness, surprise.*

Noi am tradus cuvintele din WordNet-Affect în limbile română și rusă și am creat o resursă lexicală aliniată engleză - română - rusă. Resursa dată este accesibilă gratuit în scopuri de cercetare.

1. Introducere

Creșterea spectaculoasă a tehnologiilor Web 2.0 permite fiecărui utilizator să participe activ la crearea de conținut web (bloguri, rețele sociale, chat-uri). Volumul de texte cu conținut emoțional bogat crește în progresie geometrică. Acest lucru face ca analiza subiectivă a textelor să fie de actualitate în mod special.

Până în prezent, analiza sentimentelor și studiile asupra afectelor cuvintelor s-au concentrat pe limba engleză. Un exemplu este sarcina SemEval-2007 de clasificare a textelor afective (Liu et al., 2003). De asemenea, cele mai multe resurse lexicale au fost create pentru limba engleză. De exemplu, SentiWordNet este o resursă lexicală pentru extragerea opiniilor care atribuie fiecărui synset al Wordnet-ului trei marcaje ale sentimentelor: pozitivitate, negativitate, obiectivitate (Esuli, Sebastiani, 2006).

Recent, cea mai mare creștere a utilizării Internetului a fost înregistrată pentru vorbitorii de limbi diferite de limba engleză: în anii 2000-2009, pentru limbile diferite de engleză, creșterea a depășit 690% în comparație cu 237% a creșterii pentru limba engleză[1].

În consecință, cantitatea informației textuale scrise în alte limbi decât engleza crește rapid ce sporește cererea pentru instrumentele de analiză automată a textelor și pentru resursele lexicale pentru alte limbi, diferite de engleză. Elaborarea instrumentelor a progresat pentru limbile vest-europene (franceză, germană) și asiatice (japoneză, chineză, arabă) (Edmonds, 2004), se simte nevoia de astfel de resurse și pentru limbile est-europene. Pentru a umple acest gol, am elaborat o resursă lexicală, în baza WordNet-Affect, prin traducerea în română și rusă, redactarea synset-urilor traduse și alinierea acestora la sursa engleză.

[1] http://www.internetworldstats.com/stats7.htm

2. WordNet-Affect

WordNet-Affect[2] este o resursă lexicală ce conţine informaţii despre emoţiile pe care le transmit cuvintele. În comparaţie cu WordNet-ul complet, WordNet-Affect este o resursă lexicală mică, dar valoroasă pentru adnotarea sa afectivă.

```
a#01943022 awed awestruck awestricken in_awe_of
```

Figura 1: Un synset din WordNet-Affect.

WordNet-Affect (Strapparava, Valitutti, 2004) a fost creat pornind de la WordNet DOMAINS (Magnini, Cavaglia, 2002). WordNet-Affect produce o ierarhie suplimentară a etichetelor domeniilor afective, independent de ierarhia domeniilor, cu care sunt adnotate synset-urile ce reprezintă concepte afective. „Cuvintele afective" sunt considerate a fi cuvintele ce au „conotaţie emoţională" (Ortony, 1987). Există cuvinte ce nu descriu direct unele emoţii (de exemplu, bucurie, tristeţe sau frică), dar, de asemenea, sunt legate de emoţii precum cuvintele ce descriu stările mintale, stările fizice sau trupeşti, trăsăturile de personalitate, comportamente, atitudini şi sentimente (cum ar fi plăcere sau durere).

Tabel 1: Seturile de date ale cuvintelor afective.

Clasele	#synset-uri	%synset-uri	#cuvinte	%cuvinte
anger	128	21.0	318	20.7
disgust	20	3.3	72	4.7
fear	83	13.5	208	13.5
joy	228	37.2	539	35.1
sadness	124	20.3	309	20.1
surprise	29	4.7	90	5.9
Total	612	100.0	1536	100.0

Colecţia de synset-uri din WordNet-Affect utilizată în lucrarea dată a fost furnizată ca o resursă pentru SemEval-2007 „Affective Text"[3]. Această sarcină a fost axată pe adnotarea textelor prin etichete afective (Strapparava, Mihalcea, 2008). Toate synset-urile din WordNet-Affect sunt adnotate utilizând şase etichete de categorii emoţionale: *joy, fear, anger, sadness, disgust, surprise* (Strapparava et al., 2006). Această alegere a celor şase emoţii rezultă din cercetarea psihologică a emoţiilor umane exprimate non-verbal (Ekman, 1992).

Datele despre resursa iniţială sunt descrise în tabelul 1. Toată resursa este prezentată în şase fişiere denumite conform celor şase emoţii. Fiecare fişier conţine o listă de synset-uri; un synset pe un singur rând. Un exemplu de synset este prezentat în figura 1.

Prima literă din rând indică partea de vorbire; aceasta este urmată de numărul synset-ului şi apoi sunt enumerate toate cuvintele din synset. Această reprezentare a fost simplă şi uşoară pentru prelucrarea ulterioară. În synset-uri iniţiale este un număr mare de combinaţii de cuvinte, sintagme şi expresii. Un exemplu poate fi văzut în figura 1: „`in_awe_of`". Aceste părţi ale synset-urilor au generat probleme în timpul traducerii.

[2] În scopuri de cercetare, WordNet-Affect este accesibil la cerere la http://wndomains.itc.it
[3] http://www.cse.unt.edu/~rada/affectivetext

3. Elaborarea WordNet-ului pentru limbile română şi rusă

WordNet-ul Român a fost creat de către Universitatea Alexandru Ioan Cuza din Iaşi în cadrul proiectului european Balkanet (Tufiş et al., 2006). După ce proiectul Balkanet s-a încheiat, Institutul de Cercetări pentru Inteligenţă Artificială, de la Academia Română a continuat să lucreze cu WordNet-ul Român şi, în prezent, acesta conţine 33151 synset-uri de substantive, 8929 synset-uri de verbe, 851 synset-uri de adjective şi 834 synset-uri de adverbe (Tufiş et al., 2008). Acesta poate fi accesat prin intermediul interfeţei online a MultiWordNet[4] în care WordNet-urile pentru câteva limbi sunt aliniate la WordNet-ul Princeton.

În lucrarea dată, mai întâi de toate am verificat synset-urile din WordNet-Affect folosind interfaţa online a MultiWordNet. Astfel, am copiat la setul nostru toate synset-urile care deja există în WordNet-ul Român şi nu am procesat aceste synset-uri suplimentar. Ca rezultat, în WordNet-ul Român au fost găsite 166 synset-uri, majoritatea dintre ele fiind pentru substantive şi verbe. Adjectivele şi adverbele sunt reprezentate mai puţin. Statistica privind synset-urile deja existente în română este prezentată în tabelul 2.

Tabel 2: Seturile de date a synset-urilor deja existente în Wordnet-ul Român.

Clasele	# synset-uri în WordNet-Affect	# synset-uri din WordNet Român	% synset-uri din WordNet Român
anger	116	35	30.1
disgust	17	7	41.1
fear	76	25	32.8
joy	209	63	30.0
sadness	98	24	24.7
surprise	26	12	46.1
Total	542	166	30.6

Pentru limba rusă situaţia este complet diferită. Au fost făcute mai multe încercări pentru a crea WordNet-ul pentru limba rusă. RussNet este un proiect pentru crearea tezaurului rusesc electronic (Azarova et al., 2002). Un proiect alternativ al versiunii ruseşti a WordNet-ului este WordNet-ul Rus (Balkova et al., 2004). Ambele proiecte sunt necomerciale. De asemenea, există două proiecte comerciale menite să elaboreze WordNet-uri pentru limba rusă: RuThes este un tezaur informaţional utilizat în UIS RUSSIA[5] şi proiectul privind WordNet-ul Rus creat de către grupul companiei Novosoft[6]. Din păcate, puţină informaţie şi încă mai puţine resurse sunt accesibile în mod gratuit.

4. Crearea WordNet–Affect pentru limbile română şi rusă

Pentru crearea acestor două seturi de date, am parcurs trei etape: (1) traducere automată; (2) înlăturarea traducerilor irelevante; (3) generarea synset-urilor pentru română şi rusă.

[4] http://multiwordnet.itc.it/english/home.php
[5] http://www.cir.ru
[6] http://research-and-development.novosoft-us.com

4.1 Traducerea automată

Traducerea a fost făcută automat utilizând dicţionare bilingve. Am folosit dicţionarul electronic Român–Englez ROMEN de la PRIMASOFT[7]. Dicţionarul constă din următoarele compartimente: Englez–Român, Român–Englez, Englez–Rus, Rus–Englez, fiecare conţinînd mai mult de 200 000 de intrări. Noi am utilizat doar compartimentele în care limba engleză este limba sursă. În dicţionare au fost combinări de cuvinte, sintagme şi expresii pe care le-am folosit în limbile ţintă. Pentru traducerea automată, dicţionarul a fost organizat într-o listă de cuvinte sursă urmate de traducerile ţintă. Un exemplu de intrare a dicţionarului este prezentat în figura 2.

```
Joy
Dicţionar general:
noun: bucurie; confort; fericire; plăcere; tihnă; veselie;
voioşie;
verb: a bucura; a înveseli;
```

Figura 2: Un exemplu de intrare a dicţionarului.

La această etapă, scopul nostru a fost să obţinem cât mai multe cuvinte afective posibile pentru analiză. Pentru aceasta, am tradus fiecare cuvânt din synset-urile WordNet-Affect. Am decis să excludem din synset-urile engleze toate combinările de cuvinte, sintagmele şi expresiile, deoarece ele nu pot fi traduse automat. Figura 3 prezintă un exemplu de synset tradus, obţinut după această etapă. După cum se vede în exemplu, pentru traducerea română, de asemenea am obţinut combinaţii de cuvinte care erau în dicţionar: „`a avea gust`”, „`a degusta (un aliment)`”.

Unele elemente ale synset-urilor nu au fost traduse. Acestea pot fi divizate în patru grupe. (1) Combinări de cuvinte, sintagme şi expresii pe care le-am înlăturat intenţionat din synset-urile engleze înainte de traducere. (2) Variaţii în ortografia aceluiaşi cuvânt; de exemplu, „`jubilance`”, „`jubilancy`” – primul cuvânt a fost tradus, al doilea nu a fost găsit în dicţionar. (3) Cuvinte care au fost formate cu ajutorul sufixelor „`ness`”, „`less`”, „`ful`” (de exemplu „`heartlessness`”); este puţin probabil ca acestea să apară în dicţionare, la fel ca şi adverbele formate cu sufixul „`ly`”. (4) Cuvinte care nu au fost traduse din cauza limitaţilor dicţionarului utilizat. În timp ce WordNet poate fi, în mod rezonabil, menţionat ca unul din cele mai mari dicţionare engleze, dicţionarul nostru bilingv este destul de modest. Tabelul 3 arată procentajul cuvintelor care nu au fost traduse. Media în procente a cuvintelor netraduse a fost 21%.

[7] http://www.primasoft.biz/romen_eng.php

Tabel 3: Numărul și procentajul cuvintelor netraduse.

Clase	# de cuvinte engleze	# de cuvinte traduse	# de cuvinte netraduse	% de cuvinte netraduse
anger	318	248	70	22.0
disgust	72	60	13	18.0
fear	208	162	47	22.5
joy	539	420	119	22.0
sadness	309	246	63	20,5
surprise	90	72	18	21.0
Total	**1536**	**1208**	**330**	**21.0**

A doua grupă de cuvinte nu prezintă o problemă, dar prima, a treia și a patra au trebuit traduse manual. Aceasta s-a făcut la etapa a treia.

```
05573914 n:
 preference =
          preferință
 penchant =
          înclinație
          slăbiciune
 predilection =
          predilecție
 taste =
          a avea gust
          a gusta, a cunoaște
          a gusta; a degusta (un aliment)
          degustare
          fărâmă, bucățică, îmbucătură (de)
          gust
          înclinație, preferință
```

Figura 3: Un exemplu de traducere a unui synset.

4.2 Înlăturarea traducerilor irelevante

Multe cuvinte din synset-urile engleze au câteva sensuri. Este evident că traducerea automată a generat toate traducerile posibile pentru toate sensurile. Noi am ales doar o singură traducere care era relevantă pentru semnificația synset-ului. Traducerea relevantă a fost selectată manual. Am înlăturat toate traducerile a căror sensuri nu aveau legătură cu emoția. De exemplu, cuvântul „`taste`” în synset-ul cu înțelesul de „`preference`” are câteva semnificații, însă doar ultima din lista traducerilor posibile avea legătură cu sensul comun al synset-ului. Exemplul este prezentat în figura 3. Astfel, noi am înlăturat toate traducerile cu excepția ultimei.

Deoarece am tradus fiecare cuvânt în mod separat, am obținut o mulțime de duplicate care la fel au trebuit eliminate. Am tras atenția, de asemenea, la corespondența părților de vorbire. În multe cazuri, a fost destul de dificil, în special pentru substantivele deja menționate, formate cu ajutorul sufixelor, de exemplu, „`plaintiveness`” sau „`uncheerfulness`”.

4.3 Generarea synset-urilor române şi ruse

Toate cuvintele dintr-un synset reprezintă un concept, un singur sens. Scopul etapei a treia a fost de a găsi traducerea adecvată, exact corespunzătoare acestui sens. La această etapă, a trebuit mai întâi să ataşăm explicaţia în engleză a fiecărui synset. Aceasta a determinat claritatea în semnificaţia synset-ului pentru traducători. După ce au fost adăugate explicaţiile la synset-uri, întregul set a fost dat la trei traducători care au lucrat independent. Sarcina lor a fost dublă: (1) să elimine traducerile care, din punctul lor de vedere, nu erau relevante semnificaţiei synset-ului descrise de explicaţia acestuia; (2) să adauge cât mai multe sinonime posibile relevante la synset-urile române şi ruse. Astfel, sarcina lor a fost să verifice echivalenţa sensurilor synset-urilor engleze, ruse şi române. De asemenea, au trebuit să traducă cuvintele care au rămas netraduse la prima etapă. Pentru traducere au fost folosite dicţionare online.

Dicţionarele bilingve române utilizate:

- http://hallo.ro,
- http://dictionar.netflash.ro,
- http://www.ectaco.co.uk/English-Romanian-Dictionary;

Dicţionar explicativ român: http://dexonline.ro/.

Dicţionarele bilingve ruse utilizate:

- http://en.bab.la,
- http://dictionary.babylon.com,
- http://russianlessons.net/dictionary/dictionary.php;

Dicţionare explicative ruse:

- http://slovo.freecopy.ru/,
- http://slovari.yandex.ru/dict/ushakov.

Această etapă a fost cea mai dificilă care a necesitat cel mai mult lucru. Multe synset-uri engleze aveau aproape aceleaşi sensuri, doar cu careva nuanţe. În unele cazuri, synset-urile conţineau cuvinte arhaice, care nu erau găsite în dicţionare. După cum s-a menţionat mai sus, am încercat să evităm combinările de cuvinte, sintagmele şi expresiile. Cu toate acestea, în unele cazuri, sensul exact al unui synset englez putea fi reprezentat doar de o singură combinaţie de cuvinte române sau ruse. În unele cazuri, chiar şi synset-ul englez era prezentat de combinări de cuvinte. De exemplu, `n#05591681 stage_fright`. Un alt exemplu conţine un cuvânt german: `n#05600844 world-weariness Weltschmerz`. În aşa cazuri, nu am obţinut traducerea corespunzătoare. În unele cazuri, câteva synset-uri engleze au fost traduse în aceleaşi cuvinte române sau ruse deoarece nu am putut reflecta în limba ţintă nuanţele sensurilor limbii sursă.

Referindu-se la problema cu sufixe, de exemplu, cuvintele „`weepiness`”, „`plaintiveness`”, „`mournfulness`”, „`ruthulness`”, cu greu pot fi găsite în dicţionare. Pentru a rezolva această problemă, am eliminat sufixele şi am căutat rădăcinile cuvintelor menţionate în dicţionare disponibile. În acest fel, am putut găsi sensul cuvintelor şi, prin adăugarea afixelor necesare, au fost create echivalentele în

română și rusă. De exemplu, pentru a găsi traducerea adecvată pentru cuvântul „mournfulness", am căutat în dicționar cuvântul „mournful". Rezultatul pentru limba română este „îndoliat" și pentru rusă „траурный". Având în vedere că cuvântul „mournfulness" este un substantiv, am transformat adjectivele obținute în substantive. Astfel, echivalentul românesc este „doliu" și în rusă - „траур".

Cu toate acestea, cele mai multe probleme au apărut la alinierea adjectivelor. De exemplu, pentru eticheta emoțională „*sadness"*, multe synset-uri adjectivale traduse în limba rusă conțin cuvintele „грустный" și „печальный". Prin urmare, pentru synset-uri adjectivale diferite obținem traduceri destul de asemănătoare.

4.4 Acordul între traducători

Având ca rezultatul traducerii seturi de sinonime, nu am putut folosi măsurile standard pentru acordul între traducători. Astfel, acordul a fost calculat după cum urmează. Dacă A este o mulțime de cuvinte selectate de către primul traducător pentru un synset și B este o mulțime de cuvinte selectate de al doilea traducător pentru același synset, acordul între traducători IntAgr este egal cu raportul dintre numărul de cuvinte în intersecția mulțimilor A și B și numărul de cuvinte în uniunea mulțimilor A și B:

$$IntAgr = (A \cap B) / (A \cup B) \qquad (1)$$

De exemplu, dacă un traducător a format un synset din trei cuvinte w_k, w_l și w_m, iar al doilea traducător a format același synset din patru cuvinte w_k, w_l, w_m și w_n, și primele trei cuvinte sunt aceleași, atunci $A=(w_k\ w_l\ w_m)$, $B=(w_k\ w_l\ w_m\ w_n)$, $A \cap B = (w_k\ w_l\ w_m)$, $A \cup B = (w_k\ w_l\ w_m\ w_n)$, numărul de cuvinte din intersecția A și B ar fi egal cu 3, numărul de cuvinte din uniunea A și B ar fi egal cu 4 și, prin urmare, acordul între traducători - ar fi 3 / 4 = 0.75.

Spre exemplu synset-ul „a#01195320 friendly" a fost tradus de primul traducător ca „prietenos prietenesc amical", de al doilea traducător ca „binevoitor amical prietenos", și de al treilea, ca „binevoitor prietenesc prietenos". Pentru primul și al doilea traducător intersecția traducerilor era de două cuvinte: „prietenos amical" și uniunea traducerii a fost patru cuvinte „prietenos prietenesc binevoitor amical". Acordul între traducători, în acest caz a fost de 2 / 4 = 0.5. Pentru al doilea și al treilea traducător, intersecția traducerilor era de două cuvinte: „prietenos binevoitor", și uniunea - patru cuvinte „prietenos prietenesc binevoitor amical". Prin urmare, acordul este același: 0.5. Pentru primul și al treilea traducător acordul, de asemenea, este același: 0.5. Toți cei trei traducători au în comun doar un singur cuvânt „prietenos" și uniunea de traduceri a constat din patru cuvinte. Astfel, acordul a fost de 1 / 4 = 0.25.

Tabelul 4 prezintă valorile medii ale acordului între traducători. Cei trei traducători sunt prezentați ca T1, T2 și T3.

Tabel 4: Acordul între traducători.

Perechile de traducători	Acordul între traducători
Date pentru limba rusă	
T1 – T2	0.57
T2 – T3	0.61
T1 – T3	0.59
Toţi	0.29
Date pentru limba română	
T1 – T2	0.58
T2 – T3	0.57
T1 – T3	0.67
Toţi	0.32

Unele synset-uri aveau acordul egal cu unu, spre exemplu, în synset-ul „a#00863650 euphoriant", toţi trei traducători au tradus cuvânt respectiv ca „euforizant". Cu toate acestea, pentru majoritatea synset-urilor, traducătorii au furnizat mai multe traduceri diferite, dar nu multe dintre aceste traduceri au fost comune pentru toţi traducătorii. În unele synset-uri traduse, nu a fost găsit nici un cuvânt comun în traducerile celor trei traducători. De exemplu, pentru synset-ul „a#00670851 gladdened exhilarated", cele trei traduceri au fost „bucurat înveselit înviorat bine_dispus", „bucuros vesel voios încântat bine_dispus" şi „bucurat voios bucuros înveselit". Nu a fost găsit nici un cuvânt comun pentru toate cele trei traduceri. Astfel, ne-am decis să formăm synset-uri din cuvintele care au apărut în cel puţin două dintre cele trei variante de traduceri. În acest fel, am format synset-uri finale. De exemplu, synset-ul „a#01195320 friendly" a fost tradus ca „prietenos prietenesc binevoitor amical", pentru că toate aceste cuvinte au apărut de cel puţin două ori în traduceri. Synset-ul „a#00670851 gladdened exhilarated" a fost tradus ca „bucurat înveselit bine_dispus bucuros voios".

Tabelul 5 conţine date privind numărul final de cuvinte în traduceri pentru fiecare din cele şase emoţii din WordNet-Affect .

Tabel 5: Seturile de date a cuvintelor afective din limba română şi rusă.

Clasele	#synset-urile	# Cuvinte ruse	# Cuvinte române
anger	116	393	330
disgust	17	73	60
fear	76	327	248
joy	209	765	641
sadness	98	437	364
surprise	26	129	87
Total	542	2199	1869

Trebuie de menţionat că în sursa WordNet-Affect englez au fost găsite unele synset-uri duplicate. Am eliminat toate aceste repetiţii şi numărul de synset-uri în sursa noastră este mai mic. În plus, au fost depistate unele diferenţe mici dintre WordNet-Affect,

MultiWordNet şi versiunea online a WordNet pentru că WordNet-Affect s-a creat în baza WordNet versiunea 1.6, MultiWordNet utilizează versiunea 2.0 a WordNet-ului şi versiunea online a WordNet-ului este 3.0. În ciuda numărului mai mic de synset-uri, numărul de cuvinte în seturile pentru limbile română şi rusă este mai mare decât în limba engleză. Acest lucru se datorează tendinţei noastre de a colecta în resursele noastre cât mai multe cuvinte posibile. Scopul nostru este de a folosi această resursă pentru metodele statistice de recunoaştere a emoţiilor în text.

5. Concluzii şi planuri de viitor

Acest articol descrie procesul de creare a resursei lexicale multilingve WordNet-Affect ce conţine partea engleză, partea rusă şi partea română aliniate la nivel de synset. WordNet-Affect este o resursă lexicală creată în baza WordNet-ului Princeton care codifică informaţiile despre emoţiile pe care le transmit cuvintele. Acesta este organizat în şase emoţii de bază: *anger, disgust, fear, joy, sadness, surprise*. WordNet-Affect este o resursă lexicală comparativ mică, dar valoroasă pentru adnotările sale afective.

În lucrare este descris procesul traducerii synset-urilor WordNet-Affect în limba română şi rusă şi creării WordNet-Affect aliniat Engleză - Română - Rusă. Resursa poate fi folosită pentru recunoaşterea automată a emoţiilor şi afectelor în text. Aceasta poate fi obţinută gratuit în scopuri de cercetare, pe http://lilu.fcim.utm.md. Resursa este încă în dezvoltare. Prima versiune bazată pe WordNet-Affect a fost lansată în august 2009; a doua, lansată în octombrie 2009, este aliniată cu WordNet-ul român. Resursa a fost deja folosită în (Sokolova, Bobicev, 2009) şi aceasta este numai una dintre multiplele posibilităţi de utilizare a seturilor de cuvinte.

În viitor plănuim să mărim resursa noastră. Recent am obţinut WordNet Domains care este distribuit cu WordNet-Affect. În resursa dată sunt incluse două versiuni ale WordNet-Affect original englez. Prima versiune a fost obţinută semiautomat şi conţine aproximativ 1 500 de synseturi. A doua versiune a fost creată după verificarea manuală a primei versiuni şi excluderea synset-urilor care aveau o legătură mai slabă cu emoţia respectivă. Versiunea aceasta conţine 914 synset-uri. Intenţia noastră este de a traduce toate 914 synset-uri şi de a alinia acestea la WordNet român. În afară de aceasta, vom crea resurse *'bag-of-words'* pentru utilizarea imediată în sisteme de recunoaştere a emoţiilor şi afectelor în text.

Referinţe bibliografice

Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., Oparin I. (2002). Russnet: Building a lexical database for the russian language. *Workshop on Wordnet Structures and Standardization and How this affect Wordnet Applications and Evaluation*, Las Palmas, pp. 60-64.

Balkova V., Suhonogov A., Yablonsky S.A. (2004). Russian WordNet. From UML-notation to Internet/Intranet Database Implementation. *Second International WordNet Conference, GWC 2004*, Brno, Czech Republic, 31-38.

Edmonds, P. (2002). Introduction to Senseval. *ELRA Newsletters*, 7(3), 337–344.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, vol. 6(3-4), 169–200.

Liu, H., Lieberman, H., Selker, T. (2003). A model of textual affect sensing using real-world knowledge. *ACM Conference on Intelligent User Interfaces*, Miami, Florida, USA, pp. 125-132.

Strapparava, C., Mihalcea, R. (2008). Learning to identify emotions in text. *ACM Symposium on Applied Computing*, Fortaleza, Brazil, 556-1560.

Strapparava, C., Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. *4th International Conference on Language Resources and Evaluation*, 1083–1086.

Strapparava, C., Valitutti, A., Stock, O. (2006). The affective weight of the lexicon. *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 474–481.

Tufiş, D., Mititelu, B., Bozianu, L., Mihaila, C. (2006). Romanian wordnet: New developments and applications. *3rd Conference of the Global WordNet Association*, Korea, 337–344.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., Ştefănescu, D. (2008). Romanian Wordnet: Current State, New Applications and Prospects. *4th Global WordNet Conference, GWC-2008*, University of Szeged, Hungary, 441-452.

Esuli, A., Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 417-422.

Magnini, B., Cavaglia G. (2002). Integrating subject field codes into Wordnet. *Second International Conference on Language Resources and Evaluation (LREC 2002)*, Athens, Greece, 1413—1418.

Ortony, A., Clore, G. L., Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, American Psychological Association, vol. 53, 751–766.

Sokolova M., Bobicev V. (2009). Classification of Emotion Words in Russian and Romanian Languages. *RANLP-2009 conference*, Borovets, Bulgaria, 415-419.

CAPITOLUL 3

APLICAȚII ALE TEHNOLOGIILOR LINGVISTICE TEXTUALE

SISTEM ÎNTREBARE-RĂSPUNS ANTRENABIL PENTRU LIMBA ROMÂNĂ

DAN ŞTEFĂNESCU, RADU ION, ALEXANDRU CEAUŞU, DAN TUFIŞ, ELENA IRIMIA, VERGINICA BARBU-MITITELU

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română

{danstef, radu, aceausu, tufis, elena, vergi}@racai.ro

Rezumat

Lucrarea prezintă un sistem întrebare-răspuns dezvoltat la *Institutul de Cercetări pentru Inteligenţă Artificială* – ICIA în cadrul unui proiect naţional şi evaluat independent în contextul competiţiei europene CLEF. Evaluarea a fost realizată în cadrul exerciţiului *ResPubliQA* pentru limba română. Este descris modul de combinare a diferiţilor factori de relevanţă pe baza cărora sistemul identifică paragrafele cele mai relevante ca răspunsuri la întrebările formulate în limbaj natural. Sistemul este disponibil on-line pe pagina de servicii web a ICIA. El este însă complet antrenabil, funcţionalitatea sa fiind independentă de registrul lingvistic ce caracterizează datele de antrenare.

1. Introducere

Cercetările privind prelucrarea automată a limbajului natural (PLN), domeniu central al inteligenţei artificiale, produc rezultate cu impact din ce în ce mai mare în societatea globalizată de fenomenul Internet. Sistemele de întrebare-răspuns (ÎR) în limbaj natural, în vogă în perioada anilor ‘70-80, au revenit în centrul cercetărilor PLN dar, de data aceasta având ca obiectiv identificarea răspunsurilor la întrebări arbitrare în spaţii de căutare incomparabil mai mari, la limită întregul web. Conţinutul informaţional al acestui spaţiu virtual este atât de mare încât se consideră că orice solicitare raţională de informaţie îşi poate găsi măcar un răspuns pe Internet. Evaluarea calităţii răspunsurilor şi respectiv asigurarea găsirii lor sunt însă probleme de cercetare, pentru care abordările tradiţionale au devenit insuficiente. Astfel, implementările motoarelor de căutare moderne recurg din ce în ce mai mult la tehnici PLN, acestea fiind utilizate în toate etapele fluxului de prelucrare, începând de la nivelul specificării întrebării şi până la extragerea fragmentului de text relevant dintr-unul sau mai multe documente. Odată cu utilizarea tehnicilor PLN au apărut şi campaniile de evaluare în domeniul regăsirii inteligente a informaţiei. Acestea constituie astăzi priorităţi ale cercetării de avangardă dedicată spaţiului digital al cunoaşterii. Ele au fost organizate mai întâi în SUA (MUC-*Message Understanding Conference*, TREC-*Text Retrieval Conference*, DUC-*Document Understanding Conference*, devenită TAC-*Text Analysis Conference*). În Europa, manifestarea similară este CLEF (Cross Language Evaluation Forum), ajunsă în anul 2009 la a 10-a ediţie. Având ca subiect al analizei în primul rând limbile Uniunii Europene, începând cu anul 2006 limbile europene „cu resurse electronice limitate” (română, bulgară, cehă, greacă, portugheză etc.) au devenit „subiecte” de concurs.

În acest context, proiectul naţional SIR-RESDEC (lansat în 2007) a răspuns unei priorităţi europene, propunându-şi realizarea unui sistem de ÎR în limbaj natural la nivelul celor mai avansate sisteme ale cercetării internaţionale. Consorţiul SIR-RESDEC format din cercetători de la ICIA, UAIC şi ICI şi-a concentrat eforturile în

direcţia realizării unor sisteme de ÎR pentru limbile română şi engleză, în domeniul legislaţiei Uniunii Europene şi respectiv în domeniul geneticii umane. Grupul de cercetare al ICIA şi-a concentrat eforturile în direcţia realizării sistemului de ÎR pentru limba română în domeniul legislativ, având la dispoziţie corpusul JRC-Acquis (Steinberger et al., 2006). În restul articolului de faţă vom descrie acest sistem.

2. Considerente preliminare

Pentru a testa performanţele sistemului, ICIA s-a înscris în anul 2009 în competiţia QA@CLEF la secţiunea *ResPubliQA*[1] (Peñas et al., 2009), urmând tradiţia participărilor la competiţiile CLEF încă din 2006. Sarcina sistemelor dezvoltate de echipele înscrise la *ResPubliQA* a fost să identifice automat paragrafe relevante pentru răspunsuri la întrebări formulate în limbaj natural, în domeniul juridic acoperit de corpusul paralel de lucru al competiţiei. Pentru prima oară, evaluările sistemelor de întrebare-răspuns în limbaj natural au putut fi comparate interlingual, întrucât întrebările de test (500) au fost aceleaşi în 8 limbi (bască, bulgară, engleză, franceză, germană, italiană, română şi spaniolă) răspunsurile trebuind a fi căutate în corpusul paralel (JRC-Acquis) al legislaţiei europene „Acquis Communautaire" disponibil în toate limbile Uniunii Europene. Alinierea la nivel de paragraf a corpusului pentru toate limbile implicate a oferit posibilitatea evaluării răspunsurilor sistemelor indiferent de limba de interogare. În plus, tot în premieră, organizatorii ResPubliQA au calculat, pentru fiecare limbă, performanţele unui sistem de regăsire documentară (RD) de ultimă generaţie, dar fără componente de PLN. S-a urmărit în acest mod evaluarea cantitativă a rolului tehnologiilor de prelucrare a limbajului natural faţă de tehnicile standard utilizate în regăsirea informaţiilor. Notând cu $A_{ÎR}$ acurateţea sistemului de ÎR (v. Secţiunea 6) şi cu A_{RD} acurateţea sistemului de RD, atunci raportul $M = \frac{A_{ÎR}}{A_{RD}}$ cuantifică meritul tehnicilor de prelucrare a limbajului natural. O cifră de merit supraunitară semnifică faptul că prelucrarea limbajului natural îmbunătăţeşte performanţa unui sistem de regăsire documentară. Deşi pare intuitiv ca pentru orice sistem de ÎR cifra de merit M să fie supraunitară, organizatorii ResPubliQA au constatat că dintre 28 de sisteme evaluate doar 14 au avut o cifră de merit supraunitară (v. Table 10 în (Peñas et al., 2009)).

Cu experienţa dobândită în competiţiile CLEF precedente (Puşcaşu et al., 2007; Tufiş et al., 2008c; Ion et al., 2009a) şi cerinţele specifice ale competiţiei din 2009, obiectivele tehnice principale au fost perfecţionarea modulului de regăsire a paragrafelor relevante şi respectiv a celui de reordonare a acestora, pe baza unei analize complexe a relevanţei paragrafelor candidat. Un impact semnificativ l-a constituit implementarea unei metode similare metodei de optimizare MERT (Och, 2003), în cadrul etapei de reordonare a paragrafelor. Am păstrat abordarea din anii precedenţi în ceea ce priveşte construcţia sistemului, însă diferitele module care îl alcătuiesc au fost implementate ca servicii şi/sau aplicaţii Web (Tufiş et al., 2008b):

- *serviciul de analiză a întrebării*[2] cu ajutorul căruia se clasifică fiecare întrebare ataşându-i-se o etichetă ce indică tipul de răspuns pe care acea întrebare îl cere;

[1] http://celct.isti.cnr.it/ResPubliQA/

[2] http://shadow.racai.ro/JRCACQCWebService/Service.asmx?WSDL

- *serviciile de generare a cererilor*[3] cu ajutorul căruia o întrebare în limbaj natural este transformată în interogări în limbaj formal compatibile cu motorul de căutare;
- *serviciul de interogare a motorului de căutare*[4] se ocupă de partea de regăsire a paragrafelor relevante pentru o interogare în limbaj formal furnizată la intrare;
- *modulul de reordonare a paragrafelor* preia rezultatele furnizate de motorul de căutare sub forma unei liste de paragrafe, calculează scoruri suplimentare de relevanţă pentru fiecare paragraf şi, în funcţie de o interpolare liniară (obţinută aplicând o optimizare de tip MERT) a acestora, asignează scoruri paragrafelor. Paragraful sau paragrafele cu scorurile cele mai ridicate sunt întoarse utilizatorilor.

În faza de indexare a corpusului JRC-Acquis, am luat în considerare doar textul propriu-zis al documentelor, acesta fiind în prealabil preprocesat cu ajutorul TTL (Ion, 2007) dezvoltat la ICIA. Textul a fost segmentat la nivel de unitate lexicală în funcţie de terminologia Eurovoc, adnotat la parte de vorbire, şi lematizat.

3. Identificarea terminologiei

Având în vedere caracterul juridic specializat al corpusului de lucru, o etapă importantă a fost identificarea şi tratarea ca unităţi lexicale a unor anumite expresii sau termeni multi-cuvânt. Acest lucru a fost realizat cu ajutorul tezaurului multilingv Eurovoc[5] (Ştefănescu and Tufiş, 2006). Descriptorii Eurovoc sunt termeni tehnici care trebuie să apară cu consecvenţă în toate documentele juridice pentru toate limbile implicate. Recunoaşterea acestora atât în faza de analiză a întrebărilor cât şi faza de preprocesare a documentelor (ce trebuie ulterior indexate) devine esenţială pentru performanţele oricărui sistem QA pe acest corpus.

În consecinţă am realizat un modul care, după preprocesarea corpusului, recunoaşte termenii Eurovoc şi generează unităţi lexicale corespunzătoare (în acest fel, unui termen multi-cuvânt îi corespunde o singură unitate lexicală). Etapa de identificare a terminologiei se desfăşoară de-a lungul a 6 etape: (i) termenii din Eurovoc sunt identificaţi în corpus, în forma lor de tezaur; (ii) pentru fiecare formă ocurenţă identificată la pasul 1 se extrage secvenţa de leme implicată şi se adaugă unui inventar; (iii) se identifică în corpus toate ocurenţele secvenţelor de leme din inventarul construit la pasul 2; (iv) termenii astfel identificaţi sunt aduşi la forma ocurenţă din corpus; (v) fiecărui termen identificat i se asignează un descriptor morfo-sintactic – cum termenii sunt de fapt grupuri nominale, descriptorul asignat termenului este acelaşi cu descriptorul centrului grupului nominal; (vi) fiecărei ocurenţe a unui termen i se asignează ca lemă descriptorul corespunzător Eurovoc. De exemplu, termenul *adunare parlamentară* apare în corpus cu formele flexionate: *adunarea parlamentară*, *adunările parlamentare*, *adunărilor parlamentare*. Toate aceste unităţi lexicale primesc ca lemă descriptorul Eurovoc *adunare parlamentară*, iar ca descriptor morfosintactic, descriptorul corespunzător centrului grupului (i.e., *adunare*, *adunarea*, *adunările*, etc.).

[3] http://shadow.racai.ro/QADWebService/Service.asmx?WSDL
[4] http://www.racai.ro/webservices/search.asmx?WSDL
[5] http://en.wikipedia.org/wiki/Eurovoc

4. Clasificarea automată a paragrafelor şi întrebărilor

Specificaţiile competiţiei au definit 5 tipuri de întrebări posibile: (i) *factoid* (factual) – întrebări care cer ca răspuns persoane, locaţii, instituţii, momente în timp, etc.; (ii) *definition* (definiţie) – întrebări care cer ca răspuns o definiţie; (iii) *procedure* (procedură) – întrebări care cer ca răspuns o procedură juridică; (iv) *reason* (motiv) – întrebări care cer ca răspuns un motiv, o cauză; (v) *purpose* (scop) – întrebări care cer ca răspuns un scop, un obiectiv. Numărul redus de clase şi faptul că răspunsul corect pentru o întrebare nu se poate întinde pe mai multe paragrafe au condus la ideea clasificării paragrafelor în funcţie de probabilitatea lor de a răspunde la un tip de întrebare sau altul. Etichetarea unui paragraf cu tipul de întrebare la care acel paragraf ar putea răspunde cel mai bine, oferă, în mod evident, posibilitatea reducerii complexităţii etapei de identificare a documentelor relevante, pe care o vom numi etapă de *regăsire documentară*. Acolo unde tipul întrebării este corect identificat, răspunsul este căutat, în principal, în paragrafele de tip identic cu cel al întrebării. Acest lucru a presupus construcţia unui modul de clasificare a întrebărilor care a fost optimizat şi antrenat pentru tipurile de întrebări asociate corpusului de tip juridic aşa cum vom arata mai jos.

Problema de clasificare a paragrafelor este similară cu cea a selecţiei propoziţiilor dintr-un text în vederea generării automate a rezumatului acelui text (Ion et al., 2009b). Se observă însă că cele două probleme diferă în ceea ce priveşte numărul de clase care trebuiesc considerate şi tipul entităţilor (pe de o parte paragrafe, iar de cealaltă, fraze) ce trebuie supuse procesului de clasificare. În cazul nostru, clasele pe care le-am considerat diferă uşor de cele furnizate de organizatorii competiţiei în specificaţiile date. Astfel, clasele *reason* şi *purpose* au fost unite, clasa obţinută având denumirea de *reason-purpose*. Motivul constă în dificultatea dezambiguizării automate între cele două clase. Pentru a îmbunătăţi precizia în faza de regăsire a paragrafelor relevante, am adăugat o altă clasă, etichetată *delete*, cu scopul de a elimina astfel, încă din faza de căutare, acele paragrafe care nu ar fi putut conţine răspunsuri corecte pentru vreo întrebare. În această categorie intră paragrafele ce conţin titluri (e.g., „Articolul 1”), părţi de tabele ale căror formatare nu s-a mai păstrat, semnături, etc.

Pentru faza de antrenare a clasificatorului am utilizat o colecţie de aproximativ 800 de paragrafe etichetate manual cu următoarele etichete: *factoid*, *definition*, *procedure*, *reason-purpose* şi *delete*, iar pentru a testa precizia, am folosit doar 89 de paragrafe. Metoda de clasificare la care am recurs este aceeaşi pe care am folosit-o şi pentru clasificarea întrebărilor în anii precedenţi: *principiul maximizării entropiei* (Ratnaparkhi, 1998). Trăsăturile pe care le-am luat în considerare s-au bazat pe cuvinte cheie, descriptori morfo-sintactici, punctuaţie, lungimea propoziţiei. Primele 5 cuvinte au fost considerate trăsături, ele având un puternic caracter de discriminare pentru clasele alese. Ca trăsături de tip morfo-sintactic, am luat în considerare descriptorul morfo-sintactic al verbului principal, o altă trăsătură fiind existenţa sau absenţa în cadrul propoziţiei a unui substantiv propriu. Alte trăsături sunt numărul de virgule din propoziţie, numărul de ghilimele (indicând numărul de citate), semnul de punctuaţie cu care se termină propoziţia. Din punct de vedere ortografic, o trăsătură importantă este numărul cuvintelor din propoziţie care încep cu majusculă. Trăsăturile legate de lungime includ numărul propoziţiilor în paragraf şi lungimea paragrafului în cuvinte.

Precizia clasificatorului pe datele de test a fost de 94% (Ion et al., 2009b; Ştefănescu, 2010). Deşi gradul de încredere statistică a evaluării preciziei este redus datorită numărului mic de 89 de paragrafe conţinute de datele de test, rezultatele finale obţinute folosind clasele asignate paragrafelor au fost mult îmbunătăţite. În condiţiile în care doar primele 50 de paragrafe întoarse de motorul de căutare au fost luate în considerare în etapele următoare ale fluxului de prelucrare, am constatat că, folosind clasificarea paragrafelor, în majoritatea cazurilor răspunsurile corecte s-au aflat în printre acestea.

Pentru clasificarea întrebărilor cele mai multe din sistemele actuale QA folosesc un modul specializat pentru a determina ce tip de răspuns ar trebui căutat în corpusurile avute la dispoziţie. Desigur, clasificarea se poate realiza în mai multe feluri, cele mai simple metode folosind reguli sub forma unor expresii regulate. În cazul nostru, am apelat din nou la clasificatorul bazat pe maximizarea entropiei. Acesta a fost antrenat pentru a identifica 8 clase: *reason-purpose*, *procedure*, *definition*, *location*, *name*, *numeric*, *temporal* şi *factoid*. Clasa *location* cuprinde întrebările care necesită ca răspuns o locaţie; clasa *name* conţine întrebările care cer ca răspuns un nume de persoană, organizaţie sau numele unei entităţi (e.g., comisie, ţară, etc.); clasa *temporal* cuprinde acele întrebări care cer drept răspuns o dată calendaristică sau un interval de timp; clasa *numeric* conţine întrebările care cer drept răspuns un număr („*Câţi membri sunt în comisia de …*"), iar clasa *factoid* conţine toate întrebările de tip factoid care nu sunt în clasele tocmai descrise. Trăsăturile pe care le-am folosit sunt următoarele: primul cuvânt de tip *WH* din întrebare, primul verb principal din întrebare, primul substantiv din întrebare, descriptorii morfo-sintactici ai tuturor substantivelor, verbelor, adjectivelor, adverbelor şi numeralelor din întrebare, ordinea de apariţie a primului verb şi a primului substantiv în cadrul întrebării analizate. Desigur, extragerea acestor trăsături survine abia după preprocesarea întrebării.

Pentru antrenare, am plecat de la exemplele furnizate de organizatori şi am construit 200 de întrebări (incluzând şi acele exemple) pentru care am ataşat manual clasa din care fac parte (Ştefănescu, 2010). Modelul a fost construit considerând doar trăsăturile care apăreau de cel puţin două ori în datele de antrenare, în final acesta având o acurateţe de 99% pe aceste date. Deşi evaluarea performanţelor folosind datele de antrenare („*biased evaluation*") trebuie evitată pentru aplicaţiile de învăţarea automată, ea ne-a permis totuşi să ajungem la o selecţie mai bună de trăsături. După cum era de aşteptat, acurateţea a scăzut pentru cele 500 de întrebări furnizate odată cu startul competiţiei, însă la numai 97.2%. Scorul nu este atât de surprinzător pe cât pare, datorită diferenţelor mari dintre trăsăturile caracteristice claselor considerate. De altfel, am ales clasele astfel încât clasificatorul să nu aibă probleme în identificarea lor corectă. Ne interesează în principal să nu avem erori de clasificare, chiar dacă rămânem cu clase nu foarte rafinate. Cu alte cuvinte, încercăm să restrângem cât mai mult spaţiul de căutare fără însă a îngădui prea multe erori de clasificare (Ştefănescu, 2010).

5. Sistemul QA

Sistemul nostru este implementat ca un flux de prelucrare realizat peste arhitectura serviciilor şi aplicaţiilor web dezvoltate la ICIA. Sistemul este optimizat pentru a maximiza un scor de relevanţă global $S(p)$ folosit la identificarea paragrafului cel mai plauzibil a constitui răspunsul adecvat la o întrebare adresată sistemului. Scorul global

$S(p)$ se calculează ca o combinaţie liniară a unor scoruri asignate paragrafelor în funcţie de criteriile de relevanţă în raport cu o întrebare furnizată de un utilizator:

$$S(p) = \sum_i \lambda_i s_i, \quad \sum_i \lambda_i = 1 \qquad (1)$$

unde s_i ($s_i \in [0,1]$) este unul din următoarele scoruri de relevanţă:

- s_1 este 1 când clasificarea întrebării corespunde cu cea a paragrafului, în caz contrar, valoarea sa fiind 0. După cum am arătat, clasele modulului de clasificare a paragrafelor (5 la număr) diferă de cele ale modulului de clasificare a întrebărilor (8) dar între ele există o corespondenţă bine definită (Ion et al., 2009b);

- s_2 este un scor de similaritate lexicală bazat pe lanţuri lexicale, coeziune lexicală şi identificarea perechii verb principal–argument; este scorul care încearcă să cuantifice gradul de similaritate lexicală dintre o întrebare şi paragrafele întoarse de motorul de căutare ca fiind relevante pentru acea întrebare. Se calculează după următoarea metodă: pentru o întrebare Q şi un paragraf candidat P, construim două liste conţinând lemele ce corespund cuvintelor conţinut din întrebare, respectiv din paragraf: LQ şi LP. Numim lemele din lista LQ *cuvinte cheie*. Pe baza lor asignăm lui P un scor de relevanţă. Astfel, s_2 se calculează ca un produs a trei alte scoruri ce caracterizează: (i) distanţa semantica dintre lemele celor două liste (*DS*), (ii) coeziunea cuvintelor cheie în paragraf (*CC*) şi (iii) identificarea în paragraf a unui posibil cuplu verb-argument extras din întrebare (*VA*). Aceste scoruri au fost descrise pe larg în (Ştefănescu, 2010) şi nu vom mai insista asupra lor. Scorul final se calculează ca produsul celor trei scoruri:

$$s_2 = DS \times CC \times VA$$

- s_3 este un scor similar cu scorul BLEU (Papineni et al., 2002) care avantajează paragrafele în care cuvintele cheie ale întrebării apar în aceeaşi ordine ca în întrebare; ca şi scorul precedent, se calculează considerând lemele cuvintelor conţinut din întrebare, respectiv paragraf. Ideea implementării acestui scor (Ion et al., 2009b) are la bază observaţia că de foarte multe ori, în realitate, formularea răspunsului corespunzător unei întrebări conţine o parte din acea întrebare. Principiul comparării n-gramelor a fost folosit în cazul de faţă pentru a evalua similaritatea dintre întrebare şi paragrafele candidat.

- s_4 şi s_5 sunt scorurile de relevanţă pentru paragraf şi document întoarse de motor.

4.1. Prelucrarea întrebării şi alegerea răspunsului

După ce sistemul primeşte la intrare o întrebare, ea este trimisă serviciului web TTL pentru a fi preprocesată. Este apelat apoi serviciul web care se ocupă de clasificarea întrebărilor pentru a obţine tipul de răspuns care trebuie căutat. În următorul pas, folosind informaţia adnotată după preprocesare, întrebarea este transformată în interogări într-un limbaj formal înţeles de motorul de căutare. Folosim 2 algoritmi pentru a genera două interogări diferite, ambele conţinând ca termen de căutare clasa întrebării. Trebuie menţionat că în faza de indexare au fost indexate odată cu paragrafele şi clasele corespunzătoare acestora, pentru ca reducerea spaţiului de căutare să se facă direct din faza de regăsire documentară.

Pentru fiecare din cele două interogări generate, motorul de căutare întoarce două liste L_1 şi L_2 conţinând fiecare 50 de paragrafe sortate după scorul de relevanţă descris de

ecuația (1). Răspunsul întors de sistem este acel paragraf care se găsește atât în L_1 cât și în L_2 și care satisface:

$$\text{argmin}_p(\text{rang}_1(p) + \text{rang}_2(p)), \quad \text{rang}_1(p) \leq K, \text{rang}_2(p) \leq K, K \leq 50 \qquad (2)$$

unde $rang_1(p)$ este poziția paragrafului p în L_1, iar $rang_2(p)$, poziția paragrafului p în L_2. Dacă nu există un paragraf care să satisfacă condițiile din (2), atunci sistemul semnalizează că nu poate găsi un răspuns la întrebarea dată, întorcând șirul de caractere NOA (*no answer* – nu există răspuns). În urma experimentelor, cele mai bune rezultate au fost obținute pentru K=3.

Sistemul dezvoltat este antrenabil, ponderile λ_i folosite în interpolarea liniară a scorurilor de relevanță fiind obținute cu ajutorul unei tehnici de optimizare similare cu metoda MERT. Metoda de antrenare pentru găsirea celui mai probabil răspuns a fost folosită pe cele 200 de întrebări utilizate și în cazul clasificatorului pentru întrebări și are următoarele etape: (i) rularea sistemului pentru cele 200 de întrebări și păstrarea primelor 50 de paragrafe întoarse de motorul de căutare pentru fiecare întrebare, ordinea paragrafelor fiind dată DOAR de scorurile de paragraf întoarse de motor (s_4); (ii) calcularea scorurilor s_i, $i = \overline{1,5}$, pentru fiecare din paragrafele extrase; (iii) pentru fiecare combinație de ponderi λ, cu $\sum_{i=1}^{5} \lambda_i = 1$ și un pas de incrementare de 10^{-2}, se calculează scorul *Mean Reciprocal Rank* (MRR) (Radev et al., 2002) pentru întreg setul de 200 de întrebări, lista paragrafelor întoarse fiind sortată după ecuația (1); (iv) reținerea combinației de ponderi λ care maximizează scorul MRR.

Cele două moduri de a genera cereri în limbaj formal conduc la rezultate diferite ale motorului de căutare. Putem considera că avem de a face cu două sisteme ale căror rezultate sunt combinate pentru a obține răspunsul final. Astfel, cele două sisteme au fost individual optimizate, în ceea ce privește ponderile λ, fără a avea în vedere, în vreun moment, posibilitatea ca un sistem să întoarcă răspunsuri NOA. După antrenarea celor două sisteme, observând că sistemul de evaluare al competiției *ResPubliQA* favorizează răspunsurile NOA față de cele incorecte, am luat decizia de a introduce ecuația (2) pentru a păstra doar acele răspunsuri pentru care avem un grad de încredere ridicat. La momentul transmiterii rezultatelor, setul de ponderi λ utilizat nu ținea cont de tipul întrebărilor furnizate la intrare. Vom arăta în secțiunea de evaluare că putem obține rezultate îmbunătățite antrenând setul de ponderi pentru fiecare clasă de întrebări.

4.2. Generarea de cereri formale pentru motorul de regăsire a documentelor

Etapa de regăsire documentară este similară cu cea descrisă în (Ion et al., 2009a), folosind și de această dată motorul de căutare Lucene (Hatcher and Gospodnetić, 2004). Cuvintele documentelor au fost filtrate în funcție de descrierea lor morfo-sintactică astfel încât să rămână pentru indexare doar cuvintele conținut. Mai mult, cuvintele au fost normalizate la forma lor lemă. Eventualele erori de dezambiguizare morfo-lexicală sau lematizare au fost luate în calcul și pentru a compensa astfel de erori (mai ales în cazul unităților lexicale absente din dicționarul sistemului TTL) am indexat și forma ocurență pe aceeași poziție cu lema. În acest fel, un termen poate fi căutat atât ca lemă, cât și în forma sa ocurență. Am procedat la construirea a doi indecși pentru colecția de documente: un index pentru paragrafe și unul pentru documente. Dată fiind o anumită cerere, motorul întoarce o listă alcătuită din paragrafele cele mai relevante în raport cu

acea cerere. Astfel, avem deja două scoruri calculate pentru un paragraf: scorul de relevanţă pentru paragraf în indexul construit pe paragrafe (s_4) şi cel de relevanţă pentru documentul din care face parte paragraful în indexul construit pe documente (s_5).

Abilitatea sistemului nostru de a întoarce pentru unele întrebări mesajul că nici un răspuns nu a fost găsit se datorează combinării rezultatelor diferite obţinute folosind cei doi algoritmi de generare a interogărilor formale către motorul de căutare.

ALGORITMUL DE GENERARE A CERERILOR FORMALE – VARIANTA 1 ia în considerare toate cuvintele conţinut ale întrebării (substantive, verbe, adjective şi adverbe) cu care construieşte o disjuncţie de termeni, ce sunt practic lemele acelor cuvinte. Singura condiţie impusă asupra includerii unui termen în cerere este ca scorul TF-IDF (Salton and Buckley, 1988) al acelui termen să fie peste un anumit prag (Ion et al., 2009b; Ştefănescu, 2010). În urma experimentelor, am stabilit ca valoarea acestui prag să fie 9. Folosirea acestui scor vizează eliminarea din interogare a termenilor comuni (cu frecvenţă de ocurenţă mare în documente) şi deci, cu capacitate discriminatorie mică.

ALGORITMUL DE GENERARE A CERERILOR FORMALE – VARIANTA 2 foloseşte, asemeni variantei 1, informaţia obţinută prin preprocesarea întrebării. Varianta precedentă a acestui algoritm, dezvoltată pentru căutările pe corpusul Wikipedia (Ion et al., 2009a) a fost extinsă şi optimizată pentru noul corpus. Ca şi în versiunea precedentă, algoritmul foloseşte verbele principale ale propoziţiei şi grupurile nominale identificate în faza de preprocesare pentru a genera cererea formală către motorul de căutare. Noutatea constă în faptul că pentru fiecare grup nominal sunt construiţi doi termeni ai cererii. (i) Primul este o expresie obţinută prin concatenarea cu spaţiu între ele a tuturor lemelor cuvintelor de tip conţinut în ordinea apariţiei lor în grupul nominal. Setăm apoi 2 parametri pentru expresia formată. Primul, pe care îl vom numi *parametru de proximitate*, se referă în principal la numărul de cuvinte ce pot fi intercalate între termenii expresiei. Valoarea sa este egală cu 1 plus numărul cuvintelor funcţionale ce se găsesc în grupul nominal. Al doilea parametru, pe care îl vom numi *parametru de amplificare*, defineşte importanţa expresiei ca termen al interogării. Dacă acest parametru este setat la o valoare *n*, acest lucru înseamnă că identificarea ulterioară a termenului echivalează cu identificarea a *n* termeni obişnuiţi. Am setat acest parametru la o valoare egală cu numărul unităţilor lexicale din expresia formată, pentru a favoriza paragrafele care conţin grupuri nominale identice cu cele din întrebare. (ii) Al doilea termen este o expresie Booleană formată doar din conjuncţia lemelor corespunzătoare cuvintelor conţinut din grupul nominal.

Observăm că primul termen este mai restrictiv pentru că impune o ordine de apariţie şi o anumită poziţionare în documente a cuvintelor din întrebare. În schimb, pentru al doilea termen condiţiile sunt relaxate. Urmărim ca atunci când cuvintele se găsesc în texte, dar nu strict în condiţiile impuse de primul termen, să putem extrage totuşi acele fragmente care le conţin. Ca şi în cazul versiunii precedente, pentru fiecare verb principal din întrebare (cu excepţia verbelor difuze semantic[6]) se generează un termen corespunzând lemei acelui verb. Operatorul boolean folosit implicit de Lucene este SAU (OR) şi astfel, cererile sunt disjuncţii logice. Pe de altă parte, există şi situaţia în care nici un paragraf nu conţine o anumită expresie. În acest caz, cu ajutorul expresiilor conjunctive

[6] Verbe cu putere discriminatorie redusă (a *fi*, a *avea*, a *putea*, etc.) întrucât utilizarea lor este comună în orice domeniu.

se încearcă identificarea celor care conţin cât mai multe cuvinte conţinut ce apar şi în întrebare.

Cele două cereri furnizate către motorul de căutare conduc la obţinerea a două seturi de paragrafe ca rezultate. Putem vorbi astfel de două sisteme, chiar dacă, în foarte multe din componentele ce le compun, aceste sisteme sunt similare.

Una din cele mai utilizate practici pentru a îmbunătăţi performanţele unui sistem QA este aceea de a îmbogăţi termenii mono-cuvânt din interogările formale cu sinonime extrase din lexicoane sau ontologii lexicale precum Pinceton WordNet (Fellbaum, 1998). Conţinând aproximativ 60.000 de mulţimi sinonimice, varianta românească a WordNet-ului (Tufiş et al., 2008a) este o excelentă resursă ce ar putea fi utilizată în acest scop. Cu toate acestea, datorită limbajului juridic specializat, caracteristic corpusului JRC-Acquis, am ales să nu facem uz de ontologia lexicală la nivelul generării interogărilor, ci într-o etapă ulterioară în care a fost necesară calcularea unui scor de similaritate lexicală dintre întrebare şi paragrafele având probabilitate ridicată de a conţine răspunsul corect.

5. *Evaluare şi concluzii*

Evaluarea acurateţei sistemelor înscrise în competiţia ResPubliQA a fost făcută de organizatori folosind un program ce implementează formula (3):

$$A = \frac{1}{n}\left(n_{RC} + n_{NOA} \times \frac{n_{RC}}{n}\right) \qquad (3)$$

unde n reprezintă numărul total de întrebări (500), n_{RC} este numărul de răspunsuri corecte, iar n_{NOA} reprezintă numărul întrebărilor la care am răspuns cu NOA. După cum se observă, formula de evaluare punctează suplimentar (termenul al doilea al formulei) capacitatea unui sistem de a discerne situaţiile în care nu are certitudinea necesară de a da un răspuns corect şi decide să emită un răspuns de tip „nu ştiu” (NOA=„no answer”). Ipoteza organizatorilor a fost că acurateţea în aceste situaţii ar fi aceeaşi cu acurateţea răspunsurilor corecte ($\frac{n_{RC}}{n}$). Cu alte cuvinte, atunci când sistemul decide că „nu ştie” răspunsul corect, el este la fel de precis ca în cazurile în care furnizează un răspuns. Din analiza efectuată după primirea rezultatelor competiţiei reiese că în cazul sistemului nostru, precizia cu care răspunsurile greşite au fost evitate prin soluţii NOA a fost chiar mai mare decât precizia răspunsurilor corecte.

Pentru evaluare, ICIA a trimis organizatorilor competiţiei două seturi de rezultate. Primul set, ICIA091RORO, conţine rezultatele obţinute prin combinarea ieşirilor celor două variante ale sistemului fără a folosi clasele întrebărilor ca termeni ai interogărilor formale, în timp ce al doilea set, ICIA092RORO, foloseşte aceste clase ca termeni. Utilizarea claselor întrebărilor ca termeni în interogările generate a condus la rezultate semnificativ îmbunătăţite datorită faptului că în faza de indexare a paragrafelor am adăugat un câmp care să conţină clasa acestora.

Faza de antrenare desfăşurată pe cele 200 de întrebări (v. formula (1)) a condus la obţinerea următoarelor ponderi λ:

Tabelul 4: Ponderile pentru cele două variante de interogare ale sistemului

	λ_1	λ_2	λ_3	λ_4	λ_5
varianta 1 (interogări generate în funcţie de scorul TF-IDF)	0,22	0,1	0,1	0,19	0,39
varianta 2 (interogări generate folosind structura de grupuri propoziţionale a întrebărilor	0,32	0,14	0,17	0,25	0,12

Fiecare din variantele sistemului a primit la intrare cele 500 de întrebări ale competiţiei, pentru ca apoi să întoarcă o lista a celor mai relevante 50 de paragrafe pentru fiecare întrebare, folosind parametrii din tabelul de mai sus. Dintre acestea, în cazul în care un paragraf a satisfăcut ecuaţia (2) acesta a fost furnizat ca răspuns. În caz contrar, răspunsul generat a fost NOA. Tabelul 2 conţine evaluările oficiale ale rezultatelor trimise de ICIA: ICIA091RORO şi ICIA092RORO. Scorul de acurateţe (formula 3) pentru sistemul ICIA092RORO a fost cel mai bun dintre toate cele 44 de seturi de răspunsuri (28 trimise de participanţi, plus încă 16 scoruri de referinţă calculate de organizatori). Scorul de acurateţe pentru sistemul ICIA092RORO a fost al 4-lea.

Tabelul 5: Rezultatele oficiale ale ICIA la competiţia ResPubliQA pe limba română

	ICIA091RORO)	**ICIA092RORO**
Întrebări la care a fost dat un răspuns	393	344
Întrebări la care NU fost dat un răspuns	107	156
Întrebări la care răspunsul a fost CORECT	237	260
Întrebări la care răspunsul a fost INCORECT	156	84
Întrebări la care răspunsul a fost NOA	107	156
scorul c@1	0,58	0,68

Interesant de menţionat faptul că evaluarea cifrei de merit $M = \frac{A_{IR}}{A_{RD}}$ (v. (Peñas, et al., 2009) şi Secţiunea 2) a pus în evidenţă mai obiectiv performanţa sistemelor înscrise în concurs în raport cu limba de întrebare şi răspuns. Cele două sisteme ale ICIA au obţinut cele mai mari cifre de merit (ICIA092RORO=1,55, respectiv ICIA091RORO=1,32). Următoarele cifre de merit au fost obţinute de sistemele de ÎR pentru limba italiană (1,24) şi limba spaniolă (1,175). În raport cu scorul de acurateţe c@1, aceste două sisteme s-au clasat pe locurile 8 şi respectiv 12. Aceste rezultate sunt foarte interesante şi ele demonstrează un lucru mai puţin evident: pentru limba engleză câştigul de acurateţe adus de tehnologiile de limbaj natural faţă de tehnologiile standard de regăsire a informaţiei este relativ mic (15%); acest lucru se explică prin faptul că limba engleză are o morfologie foarte simplă dar şi prin faptul că cele mai avansate tehnici de regăsire documentară au fost dezvoltate, testate şi optimizate pe documente scrise în limba engleză. Explicaţia se confirmă şi prin faptul că analiza scorurilor de acurateţe ale sistemelor pentru diferite limbi cu morfologie semnificativă (germană, franceză, spaniolă, italiană, bulgară) au fost mai mici decât scorurile de acurateţe pentru limba engleză, deşi unele dintre sistemele pentru limba engleză au fost elaborate de aceleaşi echipe care au implementat sistemele pentru limbile lor naţionale. De asemenea, ierarhizarea în funcţie de cifra de merit M, plasează cel mai bun sistem de ÎR pentru limba engleză[7] pe locul al 6-lea, în timp ce sistemul pentru limba germană, cu al 17-lea scor c@1, are cifra de merit 1,158 care-l plasează pe poziţia a 5-a.

[7] Sistemul uned092 a obţinut al doilea scor de acurateţe (c@1) 0,61, în timp ce cifra sa de merit (M) a fost de 1,151.

Am testat şi ipoteza conform căreia antrenarea ponderilor λ în funcţie de clasa întrebării va conduce la îmbunătăţirea rezultatelor. Am folosit încă odată cele 200 de întrebări pentru antrenare, partiţionate în funcţie de clasă, şi astfel am obţinut rezultatele din tabelul 3. Scorurile c@1 şi M s-au îmbunătăţit cu peste 3% faţă de cel mai bun rezultat precedent.

Tabelul 6: Ponderile rezultate după antrenarea pe fiecare clasă

		λ_1	λ_2	λ_3	λ_4	λ_5
varianta cu interogări generate în funcţie de scorul TFIDF	Factoid	0,1	0	0,2	**0,4**	0,3
	Definition	0,2	0,15	0,05	0,15	**0,45**
	Reason-Purpose	0,1	0	0,15	0,3	**0,45**
	Procedure	0,1	0	0,15	0,15	**0,6**
varianta cu interogări generate folosind structura de grupuri propoziţionale a întrebărilor	Factoid	0,15	0	**0,3**	**0,3**	0,25
	Definition	0,05	**0,5**	0,15	0,1	0,2
	Reason-Purpose	0,2	0	**0,4**	0,2	0,2
	Procedure	0,15	0,1	0,25	0,2	**0,3**

În final, trebuie să menţionăm că sistemul nostru este disponibil on-line[8], sub forma unei aplicaţii web ce apelează, pentru fiecare interogare, diferitele servicii web ce intră în arhitectura sistemului. În viitorul apropiat intenţionăm să ne concentrăm eforturile în direcţia dezvoltării unor capabilităţi cros-linguale care să dea posibilitatea utilizatorilor să interogheze sistemul în limbaj natural fie în limba română, fie în engleză, iar sistemul să întoarcă rezultatele în oricare din aceste limbi. De altfel, secţiunea în limba engleză a corpusului JRC-Acquis a fost deja preprocesată şi avem în vedere folosirea aplicaţiilor de traducere automată dezvoltate la ICIA (Ceauşu, 2009; Irimia, 2009) pentru a traduce fie întrebările în limbaj natural, fie cererile formale către motorul de căutare.

Mulţumiri. Activitatea de cercetare descrisă a fost sprijinită de CNMP – MECT prin proiectul naţional PNII SIR-RESDEC, D11007/18.09.2007.

Referinţe bibliografice

Ceauşu, A. (2009). *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă*, Bucureşti, România: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*, MIT Press.

Hatcher, E. and Gospodnetić, O. (2004) *Lucene in Action*.

Ion, R. (2007). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Ion, R., Ştefănescu, D., Ceauşu, A. and Tufiş, D. (2009a). *RACAI's QA System at the Romanian-Romanian QA@CLEF2008 Main Task*, *Lecture Notes in Computer Science*, vol. 5706, September, p. 393–400.

[8] http://www2.racai.ro/sir-resdec/

Ion, R., Ştefănescu, D., Ceauşu, A., Tufiş, D., Irimia, E. and Barbu-Mititelu, V. (2009b). *A Trainable Multi-factored QA System*, Proceedings of CLEF2009 Workshop, Corfu, Greece.

Irimia, E. (2009). *Metode de traducere autmată prin analogie. Aplicaţii pentru limbile română şi engleză*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Och, F.J. (2003). *Minimum Error Rate Training in statistical machine translation*, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan, 160-167.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation*, Proceedings of the ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, USA, 311-318.

Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, Á., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N. and Osenova, P. (2009). *Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation*, Proceedings of the QA@CLEF workshop, Sept. 30 - Oct. 3.

Puşcaşu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Ion, R., Orăşan, C., Dornescu, I., Moruz, A. and Cristea, D. (2007). *Cross-Lingual Romanian to English Question Answering at CLEF 2006*, Lecture Notes in Computer Science, Available: ISBN: 978-3-540-74998-1.

Radev, D.R., Qi, H., Wu, H. and Fan, W. (2002). *Evaluating Web-based Question Answering Systems*, Demo section, LREC 2002, Las Palmas, Spain.

Ratnaparkhi, A. (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Philadelphia, PA, USA: PhD Thesis, University of Pennsylvania.

Salton, G. and Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*, Information Processing and Management, pp. 513-523.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). *The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages*.

Ştefănescu, D. (2010). *Extragere inteligentă de informaţii din corpusuri multilingve*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Ştefănescu, D. and Tufiş, D. (2006). *Aligning Multilingual Thesauri*, Proceedings of The 5th Language Resources and Evaluation Conference (LREC), Genoa, Italy.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A. and Ştefănescu, D. (2008a). *Romanian Wordnet: Current State, New Applications and Prospects*, Proceedings of the 4th Global WordNet Conference (GWC-2008), Szeged, Hungary, 441-452.

Tufiş, D., Ion, R., Ceauşu, A. and Ştefănescu, D. (2008b). *RACAI's Linguistic Web Services*, Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco.

Tufiş, D., Ştefănescu, D., Ion, R. and Ceauşu, A. (2008c). *RACAI's Question Answering System at QA@CLEF 2007*, Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), pp. 3284-3291.

UN SISTEM DE TRADUCERE AUTOMATĂ ROMÂNĂ - FRANCEZĂ

MIRABELA NAVLEA, AMALIA TODIRAŞCU

LiLPa, Université de Strasbourg, Strasbourg – France, 22 rue René Descartes, BP 80010, 67084 Strasbourg cedex France;

mirabela.navlea@yahoo.fr, todiras@unistra.fr

Rezumat

Această lucrare prezintă un proiect de cercetare având ca obiectiv adaptarea unui sistem de traducere automată pentru perechea de limbi română - franceză. Pentru traducere, se adoptă o metodă statistică, factorizată, care combină mai multe categorii de informaţii lingvistice (categorie lexicală, proprietăţi morfologice şi sintactice). Resursele pe care le foloseşte sistemul sunt corpusuri paralele adnotate şi aliniate propoziţional şi lexical. Lucrarea prezintă corpusurile aliniate, precum şi analiza lingvistică a erorilor de aliniere lexicală identificate în corpusurile paralele aliniate. Pe baza erorilor datorate diferenţelor morfologice şi sintactice dintre română şi franceză, sunt propuse reguli euristice pentru a ameliora rezultatele alinierii lexicale.

1. Introducere

Obiectivul proiectului nostru este construirea de resurse lingvistice pentru un sistem de traducere automată statistică factorizată român - francez. Astfel, este studiată influenţa mai multor categorii de informaţii lingvistice (categorie lexicală, proprietăţi morfo-sintactice) asociate unităţii lexicale, asupra calităţii traducerilor furnizate de un asemenea sistem.

Proiectul nostru adoptă metodologia propusă în *SEE-ERA.net* (Tufiş et al., 2008). Obiectivul acestui proiect era dezvoltarea de sisteme de traducere automată statistică factorizată pentru limbi slave şi balcanice (bulgară, greacă, română, sârbă, slovenă), considerând engleza ca limbă sursă sau ţintă. Aceste sisteme utilizează corpusuri paralele adnotate şi aliniate propoziţional şi lexical.

Creşterea considerabilă a numărului documentelor disponibile în diferite limbi impune utilizarea de noi metode pentru manipularea bazelor de date voluminoase de documente multilingve, pentru crearea de situri web şi aplicaţii multilingve, pentru ameliorarea motoarelor de căutare. Unele dintre aceste metode folosesc tehnici de traducere automată care trebuie să fie adaptate pentru diferite limbi, având în vedere faptul că majoritatea sistemelor de traducere automată consideră engleza ca limbă sursă sau ţintă.

Engleza are o morfologie simplă, în raport cu alte limbi caracterizate de o morfologie flexionară complexă, cum este cazul limbilor slave şi balcanice. Pentru acestea, construirea de resurse lingvistice (dicţionare, gramatici) necesită timp şi costuri materiale şi umane importante. Cazul resurselor lingvistice româneşti este reprezentativ. Notăm existenţa mai multor dicţionare, adnotatoare şi corpusuri adnotate disponibile în limba română, dar majoritatea sistemelor de traducere automată propun româna şi engleza (Marcu & Munteanu (2005); Irimia (2008); Ceauşu (2009); sistemul *Google*

Translate[1]). La ora actuală, numai *Google Translate* tratează perechea de limbi română - franceză.

Mai mult, sistemele de traducere automată produc un număr important de erori de traducere datorate lipsei de resurse lingvistice performante pentru diferite perechi de limbi. Crearea de resurse lingvistice este o sarcină dificilă condiţionată de numeroasele fenomene lingvistice care se produc într-o limbă dată (ambiguităţi, lipsa echivalenţilor de traducere de la o limbă la alta, frazeologie). Grass (2009) identifică treisprezece tipuri de erori frecvente generate de sistemele de traducere automată: polisemia şi omonimia, ambiguitatea (sintactică, referenţială), termenii vagi (*fuzzy hedges*), expresiile şi metaforele, neologismele, substantivele proprii, cuvintele de origine străină, împrumuturile şi calcurile, separatorii, siglele şi acronimele, transpoziţia. Probleme specifice apar în cazul limbilor cu morfologie flexionară complexă, cum sunt româna şi franceza. Astfel, numărul ridicat de forme flexionare ale cuvintelor măreşte numărul de ipoteze de traducere. Pentru evitarea erorilor de acest fel, sistemele lingvistice de traducere automată se concentrează pe crearea de resurse lingvistice complexe, ca : gramatici, dicţionare, baze terminologice sau de cunoştinţe.

Dacă sistemele lingvistice (*Systran*[2]) obţin astfel rezultate de traducere performante, alte sisteme furnizează rezultate comparabile bazându-se pe tehnici statistice factorizate (*EuroMatrix*[3], 2009), ce exploatează corpusuri paralele adnotate şi aliniate. Sistemele statistice factorizate extind metodele bazate pe segmente de traducere (Koehn, Och & Marcu, 2003), ce utilizează doar forma de ocurenţă a cuvintelor, prin exploatarea mai multor factori lingvistici asociaţi unităţii lexicale, ca: leme, descrieri morfo-sintactice, informaţii sintactice etc. Aşadar, sistemele statistice factorizate sunt modulare: diferiţi factori lingvistici pot fi utilizaţi în procesul de traducere. Koehn şi Hoang (2007) utilizează proprietăţile morfo-sintactice, Avramidis şi Koehn (2008) exploatează informaţia sintactică pentru a ameliora calitatea rezultatelor de traducere.

Astfel, pentru sistemul prezentat, se adoptă o metodă de traducere statistică factorizată, adaptând însă resursele utilizate pentru perechea de limbi studiate. Corpusurile paralele disponibile pentru română şi franceză, aliniate propoziţional, adnotate şi lematizate vor fi aliniate la nivel lexical Se va studia influenţa factorilor lingvistici utilizaţi asupra calităţii traducerilor română - franceză.

În secţiunea următoare, sunt prezentate arhitectura sistemului de traducere automată statistică factorizată şi informaţiile lingvistice care sunt exploatate pentru a construi modele de limbă şi de traducere. Corpusurile utilizate sunt prezentate în secţiunea 3. Diferenţele dintre română şi franceză sunt discutate în secţiunea 4. Procesul de aliniere şi erorile reperate în corpusul aliniat lexical sunt descrise în secţiunea 5. Concluziile şi perspectivele acestui studiu sunt prezentate în secţiunea 6.

2. Metodologia adoptată

Proiectul nostru vizează dezvoltarea unui sistem de traducere automată statistică factorizată pentru română şi franceză. Acest sistem a fost iniţial implementat pentru

[1] http://translate.google.com/
[2] http://www.systransoft.com/
[3] http://www.euromatrix.net/

română şi engleză[4] (Ceauşu, 2009). Sistemul utilizează un corpus paralel adnotat aliniat propoziţional şi lexical, şi aplică diferiţi factori lingvistici: forme de ocurenţă ale cuvintelor, leme, etichete morfo-sintactice (setul de etichete MSD propus de proiectul Multext[5] pentru franceză (Ide & Véronis, 1994) şi română (Tufiş & Barbu, 1997)), grupuri nominale sau verbale nerecursive (chunk-uri), colocaţii.

Sistemul îşi dovedeşte eficienţa (Ceauşu, 2009), în principal, pentru traduceri din domeniul juridic - administrativ. Sistemul utilizează decodorul *MOSES* (Koehn et al., 2007), cu diferite configuraţii ale parametrilor lingvistici optimizaţi (leme, MSD), stabilite în funcţie de sensul procesului de traducere.

Pentru a adapta *MOSES* la o nouă pereche de limbi, este necesară construirea unui model de limbă, utilizând un corpus monolingv adnotat în limba ţintă, şi a unui model de traducere factorizat, utilizând un corpus paralel adnotat şi aliniat propoziţional şi lexical. Apoi, decodorul *MOSES* caută traducerea cea mai probabilă utilizând modelele de limbă şi de traducere construite.

Cum sistemul nu a fost adaptat pentru română şi franceză, studiul nostru vizează construirea resurselor lingvistice necesare pentru obţinerea modelelor de limbă şi de traducere pentru aceste limbi. Astfel, se iau în considerare următoarele etape:

1) constituirea corpusurilor paralele;
2) procesarea corpusurilor (segmentare lexicală, lematizare, adnotare morfo-sintactică şi la nivelul grupurilor sintactice nerecursive);
3) alinierea propoziţională şi lexicală a corpusurilor paralele;
4) corectarea erorilor de aliniere, în urma unei analize detaliate a acestor erori. Aceste date sunt utilizate pentru a reantrena modulul de aliniere;
5) construirea modelelor de limbă în limba ţintă, utilizând corpusuri monolingve adnotate, şi a modelelor de traducere factorizate, utilizând corpusuri paralele adnotate şi aliniate, pentru română şi franceză;
6) configurarea sistemului cu factorii lingvistici cei mai relevanţi (optimizare);
7) analiza lingvistică a erorilor de traducere, reconfigurarea sistemului prin reluarea etapei 5 şi optimizarea sistemului.

În secţiunile următoare, sunt prezentate corpusurile monolingve şi paralele utilizate. Sunt prezentate, de asemenea, etapele de procesare, procesul de aliniere lexicală şi un ansamblu de erori de aliniere reperate în corpusul aliniat român - francez.

3. Corpusurile

Se utilizează un corpus paralel român - francez extras din *JRC-Acquis* (Steinberger et al., 2006). *JRC-Acquis* se bazează pe corpusul paralel multilingv *Acquis Communautaire* (legislaţia UE, 1950-prezent) şi este disponibil pentru 231 de perechi de limbi obţinute din 22 de limbi oficiale ale UE. *JRC-Acquis* este aliniat la nivel de paragraf şi disponibil gratuit în format XML. În proiectul nostru, se foloseşte un subset

[4] http://www.racai.ro/webservices
[5] http://aune.lpl.univ-aix.fr/projects/multext/

de 228 174 de perechi de fraze aliniate 1:1 din *JRC-Acquis* (5 357 017 de cuvinte în română, 5 828 169 de cuvinte în franceză), selectate din ansamblul de documente comune în română şi franceză.

Se foloseşte, de asemenea, un corpus paralel român - francez aliniat propoziţional, extras din *DGT Translation Memory (DGT-TM)*[6]. Acest corpus se bazează tot pe *Acquis Communautaire*, dar are avantajul că majoritatea frazelor aliniate este corectată manual. *DGT-TM* în română şi franceză conţine 490 962 de perechi de fraze aliniate (9 142 291 de cuvinte în română şi 9 953 360 de cuvinte în franceză). Acest corpus e disponibil gratuit în format TMX.

Cum *JRC-Acquis* şi *DGT-TM* sunt corpusuri specializate juridic - administrativ, se utilizează şi alte corpusuri paralele, constituite folosind resurse disponibile pe Internet, pentru a testa sistemul de traducere automată şi pentru alte domenii (politică, aviaţie). Pentru colectarea corpusului paralel român - francez, au fost luate în considerare criterii ca: disponibilitatea textelor bilingve, fiabilitatea sursei textelor, calitatea traducerilor, precum şi domeniul. Astfel, corpusul paralel a fost selectat şi organizat manual ţinându-se cont de criteriile menţionate şi specificându-se pentru fiecare text sursa indicată prin URL, autorul, data. Datele colectate au fost curăţate prin eliminarea elementelor netextuale, ca: imagini, tabele, note de subsol etc. Pentru rezolvarea problemei lipsei diacriticelor pentru majoritatea textelor româneşti colectate de pe Internet, s-a folosit sistemul de recuperare a diacriticelor, *Diac+* (Tufiş & Ceauşu, 2008). Astfel, textele bilingve disponibile în română şi franceză au fost colectate folosind mai multe situri web: documente de pe situl Parlamentului European (263 788 de cuvinte), siturile companiilor aeriene româneşti (63 353 de cuvinte).

Pentru evaluarea sistemului nostru, se utilizează un corpus român - francez aliniat propoziţional şi lexical, constituit din 1000 de fraze (Todiraşcu et al., 2008). Acesta a fost obţinut printr-un proces de derivare (Tufiş & Koeva, 2007), folosindu-se două corpusuri paralele extrase din *JRC-Acquis* (englez - român şi englez - francez) şi aliniate lexical în mod automat. Corpusul aliniat lexical român - francez a fost corectat manual.

Pentru construirea modelelor de limbă, se utilizează următoarele corpusuri monolingve pentru limba română[7]:

(i) partea românească a corpusului jurnalistic paralel *NAACL* (800 000 de cuvinte) (Martin, Mihalcea & Pedersen, 2005);

(ii) corpusul *LT4eL* (alcătuit din manuale de utilizare, 600 000 de cuvinte) (Trandabăţ et al., 2006);

(iii) corpusul jurnalistic *RoCo* (7,5 milioane de cuvinte) (Tufiş & Irimia, 2006).

Pentru limba franceză, se folosesc corpusurile următoare:

(i) un corpus juridic - administrativ selectat din *JRC-Acquis* (498 788 de cuvinte);

(ii) un corpus jurnalistic selectat din *Le Monde* (1980-1988) (488 543 de cuvinte).

Corpusurile se procesează prin aplicarea tagger-ului *TTL*[8] (Ion, 2007), disponibil pentru română şi franceză ca serviciu web. *TTL* segmentează lexical, lematizează şi adnotează

[6] http://langtech.jrc.it/DGT-TM.html
[7] Corpusuri disponibile la cerere de la autori.

textul cu descrieri morfo-sintactice şi la nivelul grupurilor nominale, prepoziţionale şi verbale nerecursive (chunk-uri). Rezultatele furnizate sunt în format XCES (Figura 1) şi conţin setul de etichete MSD propus de proiectul Multext pentru franceză (Ide & Véronis 1994) şi română (Tufiş & Barbu, 1997).

```
<seg lang="fr"><s id="ttlfr.3">
<w lemma="voir" ana="Vmps-s">vu</w>
<w lemma="le" ana="Da-fs" chunk="Np#1">la</w>
<w lemma="proposition" ana="Ncfs" chunk="Np#1">proposition</w>
<w lemma="de" ana="Spd" chunk="Pp#1">de</w>
<w lemma="le" ana="Da-fs" chunk="Pp#1,Np#2">la</w>
<w lemma="commission"ana="Ncfs"chunk="Pp#1,Np#2">Commission
</w>
<c>;</c>
</s></seg>
```

Figura 1: Un exemplu de rezultate furnizate de *TTL* pentru franceză.

4. Construirea corpusurilor aliniate

După cum este menţionat şi în secţiunea anterioară, se utilizează corpusuri paralele adnotate şi aliniate propoziţional şi lexical, pentru construirea modelelor de traducere factorizate. Calitatea corpusurilor aliniate este esenţială pentru a obţine rezultate de traducere performante. Cum nu dispunem de resurse pentru alinierea lexicală română - franceză, se utilizează corpusul aliniat propoziţional (Ceauşu et al., 2006), care este, de asemenea, lematizat, etichetat şi adnotat la nivelul grupurilor nerecursive.

Se pregăteşte corpusul în formatul de intrare cerut de sistemul statistic *GIZA++*[9] (Och & Ney, 2003), furnizându-se, de asemenea, lema şi informaţia de categorie lexicală pentru dezambiguizarea morfologică a lemei. Se efectuează o aliniere lexicală bidirecţională, apoi se obţine intersecţia aliniamentelor.

Se filtrează lista echivalenţilor de traducere astfel obţinuţi, cu o listă de cuvinte ce au ortografie asemănătoare şi un sens comun (numite *cognates* în engleză) pentru română şi franceză. Pentru identificarea acestor cuvinte, se foloseşte un algoritm care calculează cea mai lungă secvenţă de caractere comună pentru două cuvinte date. Dacă lungimea celei mai lungi secvenţe comune de caractere este cel puţin 70% din lungimea celui mai scurt cuvânt, atunci perechea de cuvinte este considerată *cognat*.

Se utilizează lista filtrată a echivalenţilor de traducere, ca punct de plecare pentru efectuarea alinierii lexicale.

Cum a fost propus în Tufiş et al. (2005), se aplică un ansamblu de euristici pentru realizarea alinierii lexicale: (i) definirea claselor de echivalenţă a categoriei lexicale (un substantiv poate fi tradus printr-un substantiv, verb sau adjectiv); (ii) alinierea substantivelor, adjectivelor, verbelor, adverbelor; (iii) alinierea grupurilor nominale, prepoziţionale sau verbale nerecursive conţinând echivalenţi de traducere; (iv) alinierea elementelor conţinute într-un astfel de grup, prin aplicarea de reguli euristice.

[8] TTL = Tokenizing, Tagging and Lemmatizing free running texts.
[9] http://fjoch.com/GIZA++.html

Pentru ameliorarea rezultatelor alinierii lexicale, au fost analizate erorile sistematice de aliniere. Astfel, a fost definit un ansamblu de reguli euristice contextuale pentru evitarea acestor erori. În secţiunea 5.2., sunt prezentate câteva dintre regulile definite.

5. Erorile de aliniere lexicală

S-a aliniat un corpus român - francez conţinând 1000 de fraze extrase din *JRC-Acquis*, aşa cum este descris în secţiunea anterioară. Au fost analizate apoi erorile sistematice de aliniere lexicală. Cele mai multe dintre acestea apar din cauza diferenţelor morfo-sintactice dintre limbi, alte erori fiind specifice domeniului: colocaţii, termeni.

5.1. Comparaţie între română şi franceză

În ciuda faptului că româna şi franceza sunt amândouă limbi romanice, fiecare posedă caracteristici morfologice şi sintactice specifice. Structurile sintactice sunt asemănătoare, dar caracteristicile morfo-sintactice prezintă diferenţe importante. Astfel, substantivele şi nominalele din română sunt caracterizate de caz (nominativ, acuzativ, genitiv, dativ, vocativ), marcat flexionar prin desinenţe specifice şi analitic, prin morfeme sau afixe proclitice (*GALR*[10], 2005), şi pot fi de genul neutru. Determinantul definit din română este întotdeauna enclitic şi fuzionează cu substantivul şi cu adjectivul antepus, în timp ce în franceză acesta este întotdeauna proclitic, formând cuvânt separat. În română, cliticele de acuzativ şi de dativ se exprimă simultan cu complementul direct sau indirect. În franceză, utilizarea pronumelui clitic exclude realizarea de suprafaţă a complementului direct sau indirect ca grup nominal. Alte diferenţe privesc structurile sintactice specifice fiecărei limbi (subordonatele relative), construcţiile verbale specifice (morfeme adiţionale pentru anumite moduri sau timpuri în română, auxiliare specifice pentru verbe de mişcare în franceză), şi morfemele diferite de la o limbă la alta (relaţia de posesie).

5.2. Analiza lingvistică a erorilor de aliniere lexicală

În această secţiune, sunt prezentate câteva dintre cele mai frecvente erori de aliniere lexicală reperate în corpusul studiat, extras din *JRC-Acquis*. Aceste erori sunt specifice domeniului juridic - administrativ al corpusului studiat, caracterizat de folosirea construcţiilor impersonale, a persoanei a-III-a singular sau plural etc. Pentru corectarea erorilor, a fost definit un ansamblu de 27 de reguli euristice morfo-sintactice contextuale. Aceste resurse se utilizează pentru efectuarea alinierii lexicale în interiorul grupurilor nerecursive (chunk-urilor). Se dau câteva exemple de erori şi de reguli, acestea fiind explicate în (Navlea & Todiraşcu, 2010).

Subordonatele relative. Una dintre erorile frecvente priveşte subordonatele relative. În subordonatele relative din română, complementul indirect este dublu exprimat, prin pronumele relativ *care* (dativ) şi formele neaccentuate ale pronumelui personal *îi, -i, le, li*. În subordonatele relative din franceză, complementul indirect se exprimă cu ajutorul prepoziţiei *à* sau *auprès de* şi al pronumelui relativ *lequel*. Din cauza acestei diferenţe, în exemplul din Figura 2, pronumele personal *li* este nealiniat cu pronumele relativ *auxquelles* (formă contractată *à + lesquelles*).

[10] Gramatica limbii române.

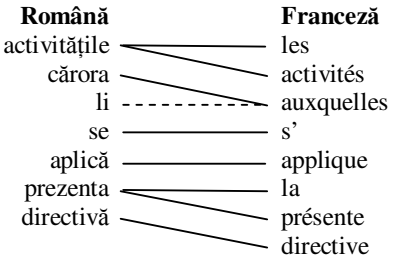


Figura 2: Eroare de aliniere lexicală privind subordonatele relative cu complement indirect.

Un exemplu de regulă euristică morfo-sintactică contextuală pentru evitarea acestui tip de erori, este dat în tabelul 1 de mai jos :

Tabel 1: Regulă euristică pentru subordonatele relative cu complement indirect

Română	**Franceză**
N + lema *care* (dativ) + *îi*|*-i*|*le*|*li* + V	N + *à*|*auprès de* + lema *lequel* + V

Destinatarul. Erori de aliniere frecvente apar şi în cazul exprimării destinatarului. În franceză, destinatarul este exprimat cu ajutorul prepoziției *à*, în timp ce în română acesta se exprimă prin dativ. Din cauza acestei diferențe, în exemplul din Figura 3, prepoziția *à* din franceză nu este aliniată cu substantivul în dativ din română.

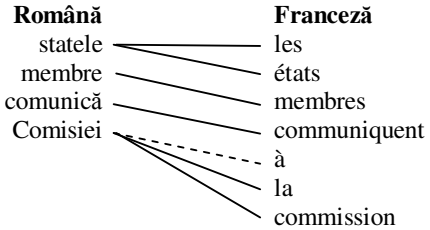


Figura 3: Eroare de aliniere lexicală privind exprimarea destinatarului.

Tabelul 2 de mai jos conține un exemplu de regulă euristică contextuală propusă pentru a evita această categorie de erori.

Tabel 2 : O regulă euristică pentru exprimarea destinatarului

Română	**Franceză**
V + N determinat definit (dativ)	V + *à* | *(au, aux)* + determinant definit + N

Colocațiile. Se corectează**,** de asemenea, erorile de aliniere lexicală privind colocațiile. Colocațiile sunt expresii polilexicale, ale căror cuvinte sunt legate prin relații lexico-sintactice (Todirașcu et al., 2008). Aceste expresii nu sunt aliniate în bloc (Figura 4).

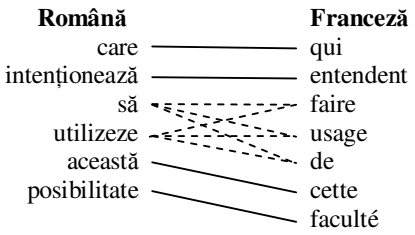


Figura 4: Eroare de aliniere lexicală privind colocaţiile.

Se efectuează alinierea colocaţiilor verbo-nominale prin utilizarea unui dicţionar multilingv de colocaţii verbo-nominale, ce conţine şi proprietăţile lor morfo-sintactice contextuale (Todiraşcu et al., 2008). Colocaţiile sunt utilizate astfel ca indici principali în procesul de aliniere lexicală. Cum dicţionarul a fost completat cu date extrase din corpusul *JRC-Acquis*, şi conţine cele mai frecvente colocaţii verbo-nominale, existente în corpusurile francez, român, englez şi german, acesta este eficient pentru alinierea colocaţiilor verbo-nominale, însă nu rezolvă problema altor clase de colocaţii (nominale, adverbo-adjectivale etc.). Pentru acestea, se vor căuta alte strategii de aliniere.

6. Concluzii

Lucrarea de faţă prezintă un proiect în desfăşurare ce urmăreşte construirea de resurse lingvistice pentru sisteme de traducere automată statistică factorizată, pentru două limbi cu morfologie flexionară complexă : româna şi franceza. Sunt prezentate resursele utilizate pentru construirea corpusurilor aliniate lexical. Sunt analizate rezultatele modulului de aliniere lexicală şi sunt propuse reguli euristice morfo-sintactice contextuale pentru ameliorarea acestor rezultate. Până în acest moment, au fost implementate regulile euristice şi evaluarea rezultatelor nu este încă terminată. În viitor, se vor construi modele de traducere factorizate pentru română şi franceză. Se vor compara rezultatele cu modele de traducere statistice pure şi se vor evalua mai multe combinaţii de factori lingvistici, pentru găsirea parametrilor optimali în ceea ce priveşte perechea de limbi studiate.

Mulţumiri. Autorii mulţumesc doamnei Rada Mihalcea (University of Texas) pentru corpusul *NAACL*, d-lui Dan Cristea (Universitatea din Iaşi, România) pentru corpusul *LT4eL* şi d-lui Dan Tufiş (Academia Română, Bucureşti) pentru corpusul *RoCo.*

Referinţe bibliografice

Avramidis, E., Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation, In *Proceedings of ACL-08: HLT,* Columbus, June 2008, pp. 763-770.

Ceauşu, A., Ştefănescu, D., Tufiş, D. (2006). Acquis Communautaire Sentence Alignment using Support Vector Machines, In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006, pp. 2134-2137.

Ceauşu, A. (2009). *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă*, Teză de doctorat, Academia Română, Bucureşti, aprilie 2009, 123 p.

Grass, T. (2009). A quoi sert encore la traduction automatique? *Les Cahiers du GEPE, Outils de traduction - outils du traducteur?*, n° 3, Strasbourg, 14 p.

Guţu Romalo, V. (coord.) (2005). *Gramatica limbii române*, *Cuvântul*, vol. I, Ed. Academiei Române, Bucureşti, 712 p.

Ide, N., Véronis, J. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, Kyoto, August 5-9, pp. 90-96.

Ion, R. (2007). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*, Teză de doctorat, Academia Română, Bucureşti, mai 2007, .

Irimia, E. (2008). Experimente de Traducere Automată Bazată pe Exemple, *Atelierul de Lucru Resurse Lingvistice Româneşti şi Instrumente pentru Prelucrarea Limbii Române*, Iaşi, 19-21 novembre 2008, pp. 131-140.

Koehn, Ph., Och, F. J., Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*, Edmonton, May-June 2003, pp. 48-54.

Koehn, Ph., Hoang H., (2007). Factored translation models, In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, June 2007, pp. 868-876.

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, Ch., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Ch., Zens, R., Dyer, Ch., Bojar, O., Constantin, A. Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007* , Prague, June 2007, pp. 177-180.

Marcu, D., Munteanu, D. S. (2005). Statistical Machine Translation: An English-Romanian Experiment, *EUROLAN* 2005.

Martin J., Mihalcea R., Pedersen T. (2005). Word Alignment for Languages with Scarce Resources. In *Proceeding of the ACL2005 Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*. June, 2005, Ann Arbor, Michigan, Association for Computational Linguistics, 65-74

Navlea, M., Todiraşcu, A. (2010). Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems**,** In *Proceedings of LREC Workshop Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages,*Valletta, May 2010, pp. 41-48.

Och, F. J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, vol. 29, n° 1, March 2003, pp. 19-51.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006, pp. 2142-2147.

Todiraşcu, A., Heid U., Ştefănescu, D., Tufiş D., Gledhill C., Weller M., Rousselot F. (2008). Vers un dictionnaire de collocations multilingue. *Cahiers de Linguistique*, vol. 33, n° 1, Louvain, août 2008, p. 161-186.

Trandabăț, D., Iftene, A., Pistol, I., Forăscu, C., Cristea, D. (2006). Resurse românești în cadrul proiectului LT4eL. În C. Forăscu, D. Tufiș, D. Cristea (eds.): Resurse lingvistice și instrumente pentru prelucrarea limbii române, Editura Universității „Al.I. Cuza" Iași, România, ISBN 978-973-703-208-9.

Tufiş, D., Barbu, A., M. (1997). A Reversible and Reusable Morpho-Lexical Description of Romanian, In Dan Tufiș and Poul Andersen (eds.), *Recent Advances in Romanian Language Technology*, pp. 83-93, Editura Academiei Române, București, 1997. ISBN 973-27-0626-0.

Tufiş, D., Ion, R., Ceauşu A., Ştefănescu, D. (2005). Combined Aligners. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 107-110, Ann Arbor, USA, Association for Computational Linguistics. ISBN 978-973-703-208-9.

Tufiş, D., Irimia, E. (2006). RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 869-872, Genoa, Italy, May 2006. ELRA - European Language Ressources Association.

Tufiş, D., Koeva, S. (2007). Ontology-supported Text Classification based on Cross-lingual Word Sense Disambiguation. In Masulli, F., Mitra, S., Pasi, G.., eds., *Applications of Fuzzy Sets Theory. 7th International Workshop on Fuzzy Logic and Applications (WILF 2007)*, volume 4578 of Lecture Notes in Artificial Intelligence, September 2007, Springer - Verlag, pp. 447-455.

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2008). Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Tadić (M.) et al., eds, In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, Dubrovnik, September 25-28, pp. 145-152.

Tufiş, D., Ceauşu, A. (2008). DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May, 2008, ELRA, ISBN 2-9517408-4-0.

SERVICII WEB INTEROPERABILE ŞI MULTILINGUALE

RADU ION, ALEXANDRU CEAUŞU, DAN ŞTEFĂNESCU, DAN TUFIŞ

Institutul de Cercetări pentru Inteligență Artificială, Academia Română

{radu, aceausu, danstef, tufis}@racai.ro

Rezumat

Problemele interoperabilității instrumentelor și resurselor lingvistice sunt preocupări majore ale cercetării actuale în domeniul prelucrării limbajelor naturale. Cele mai noi tehnologii, bazate pe arhitecturi orientate pe servicii web, constituie un pas important în direcția asigurării interoperabilității. Atunci când serviciile web sunt independente de limbă sau facil adaptabile la diverse limbi criteriile de interoperabilitate și multilingualitate sunt în mare măsură satisfăcute. Orchestrarea diverselor servicii se face de obicei prin alegerea unui format comun al datelor de intrare/ieșire. În acest sens, au apărut deja o serie de platforme de integrare a resurselor și instrumentelor lingvistice dezvoltate și rezidente în locații diferite, implementate în limbaje de programare și sub sisteme de operare variate. În această lucrare prezentăm câteva dintre serviciile web ale Institutului de Cercetări pentru Inteligență Artificială (ICIA) adaptate la platforma de servicii web WebLicht. WebLicht reprezintă un mediu în care serviciile web înregistrate pot interacționa unele cu altele datorită conversiei parametrilor de intrare/ieșire la formatul TCF (engl. Text Corpus Format).

1. Introducere

Tehnologiile Web Service, un palier fundamental în filosofia noilor generații ale web-ului, înlesnesc utilizatorilor dezvoltarea de aplicații ce integrează diverse module, implementate de autori diferiți, în limbaje diferite, și chiar aflate pe alte mașini decât cea locală. Conceptul de arhitectură orientată spre servicii (SOA) a apărut dintr-o nevoie imperioasă de a structura și standardiza colecția eterogenă și vastă de instrumente și documente care există în spațiul virtual al intra- și inter-rețelelor informatice locale, regionale sau mondiale. Derivat din conceptul SOA, dar mai general, este conceptul de infrastructură de servicii web care oferă răspunsuri la construcția de fluxuri de prelucrare pe baza serviciilor web dar și la identificarea serviciilor relevante pentru o anumită aplicație, a datelor necesare diferitelor servicii, a publicării de resurse și instrumente de prelucrare a acestor resurse, a documentării lor etc. O astfel de infrastructură este în curs de construcție în cadrul proiectului european CLARIN[1].

CLARIN își propune să reunească aplicații și resurse din domeniul Prelucrării Automate a Limbajului Natural (PLN) și să le prezinte sub o formă în care nespecialiștii în prelucrarea limbajului natural (PLN) să le poată identifica și utiliza cât mai ușor în activitățile pe care le întreprind. Altfel spus, CLARIN urmărește realizarea unei infrastructuri europene a tehnologiilor limbajului natural (scris și vorbit) astfel încât acestea să fie *integrate* (aplicațiile și resursele se află în centre specializate care utilizează servere dedicate interconectate prin rețele de tip GRID), *interoperabile*

[1] http://www.clarin.eu

(resursele şi serviciile vor fi descrise cu limbajele proprii Web-ului semantic (engl. Semantic Web) pentru a depăşi barierele puse de diversitatea formatelor de intrare/ieşire), *stabile* (garantarea funcţionării neîntrerupte şi existenţa personalului de suport tehnic), *persistente* (garantarea funcţionării continue pe o perioadă lungă de timp – cel puţin pe durata proiectului), *accesibile* (resursele şi aplicaţiile sunt accesibile online pe Web) şi *extensibile* (întreg mediul trebuie să incorporeze uşor noi aplicaţii şi resurse). CLARIN reuneşte şi armonizează la nivel european o multitudine de proiecte naţionale. Astfel de proiecte sunt, printre altele, proiectul german D-SPIN şi proiectul românesc CLARIN-RO.

D-SPIN[2] (Hinrichs et al., 2008) a implementat o platformă de servicii web care să asigure interoperabilitatea unor aplicaţii de PLN dezvoltate independent. Metoda de interconectare a fost aceea a expunerii acestor aplicaţii ca servicii web de tip REST[3] (şi nu SOAP[4] pentru că s-a considerat că volumul de lucru suplimentar necesar împachetării/despachetării formatului SOAP va introduce un timp suplimentar şi inutil de adaptare pentru noile aplicaţii care doresc înregistrarea). Această platformă se numeşte **WebLicht**[5] şi a fost dezvoltată în colaborare de universităţile din Tübingen, Stuttgart, Leipzig şi Berlin. În prezent conţine peste 70 de aplicaţii de PLN (pentru limbile germană, engleză, franceză, română, spaniolă, italiană şi finlandeză) expuse ca servicii web şi care *sunt compatibile între ele ca formate de intrare/ieşire*. Acest lucru permite în mod evident posibilitatea extraordinară a compunerii de operaţii simple pentru obţinerea unor rezultate care altfel ar fi fost imposibil de obţinut – de exemplu, combinarea adnotării cu etichete morfosintactice a unui text din două surse (programe) diferite. Dintre operaţiile ce se pot efectua asupra unui text putem menţiona: segmentare lexicală, adnotare cu etichete morfo-sintactice, lematizare, analiză sintactică etc.

Formatul parametrilor de intrare/ieşire este unul XML care a fost definit special pentru a facilita integrarea aplicaţiilor care prelucrează texte. Acesta se numeşte Text Corpus Format (TCF, (Schmid, 2009)) şi reprezintă o stivă de adnotări care sunt produse de operaţii elementare de procesare a textelor (de exemplu, segmentarea la nivel de unitate lexicală este o adnotare necesară operaţiei de etichetare morfo-sintactică). Astfel, fiecare serviciu web din WebLicht aşteaptă la intrare un fişier XML care conţine textul prelucrat până la un anumit nivel şi întoarce acelaşi fişier XML la care adaugă un nou nivel care conţine adnotările pe care le produce (niciunei operaţii nu-i este permis să şteargă vreun nivel de adnotare). Menţionăm că platforma Weblicht este una dintre componentele ECD (European Clarin Demonstrator) prototipul funcţional al etapei de specificare (preparatory phase) a proiectului CLARIN, prototip a cărui implementare este coordonată de ICIA.

În cele ce urmează, vom descrie procesul de adaptare, în cadrul proiectului RO-CLARIN, a câtorva dintre serviciile web ale ICIA la platforma WebLicht. O serie de alte servicii web proprii sau în curs de implementare la ICIA vor fi adăugate în viitorul

[2] http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml
[3] Representational State Transfer, un tip de comunicare de tip client-server.
[4] Ajuns la versiunea 1.2, SOAP (http://www.w3.org/TR/soap12-part0/) este un protocol descris în XML cu care diverse aplicaţii distribuite pe calculatoare conectate în reţea, pot comunica prin schimb de mesaje împachetate în formatul SOAP.
[5] http://weblicht.sfs.uni-tuebingen.de:8080/WebLicht1/

apropiat platformei WebLicht. De asemenea vom exemplifica avantaje ale adoptării de standarde pentru uşurinţa adaptării unor aplicaţii diverse (expuse ca servicii web) la rezolvarea unor probleme care altfel nu ar fi putut fi abordate. În acest sens, vom arăta cum anume putem combina rezultatele mai multor programe de etichetare morfo-sintactică operând pe aceeaşi segmentare lexicală cu metoda descrisă în (Tufiş, 1999) care opera cu acelaşi program de etichetare morfo-sintactică antrenat însă pe diverse corpusuri.

2. TTL în platforma WebLicht

TTL[6] este un modul Perl care oferă următoarele adnotări ale textelor în limbile română, engleză şi franceză: segmentare la nivel de frază şi unitate lexicală, adnotare cu etichete morfo-sintactice (engl. POS tagging), lematizare, recunoaşterea entităţilor denumite (engl. named entity recognition) şi analiză sintactică de suprafaţă (engl. chunking). A ajuns la versiunea 7.9 şi a fost implementat deja ca serviciu web (Tufiş et al., 2007).

Pentru adaptarea TTL la WebLicht a fost nevoie de următoarele modificări asupra implementării anterioare ca serviciu web:

- renunţarea la suportul pentru SOAP întrucât WebLicht specifică comunicarea de tip REST ca fiind standardul acceptat. În acest caz, un client care doreşte să folosească serviciul web va trimite o cerere HTTP POST către serverul de web care găzduieşte serviciul comunicându-i acestuia un şir de caractere codificat UTF-8 care reprezintă un document XML TCF bine format. Serviciul web procesează documentul şi îl returnează clientului cu stratul de adnotări suplimentar pe care l-a produs.
- renunţarea la serverul de web implementat de pachetul Perl **SOAP::Lite** şi adoptarea serverului de web **Apache**[7] ca gazdă pentru TTL. În acest fel am rezolvat o problemă gravă a serverului de web din SOAP::Lite care nu accepta decât o singură conexiune la serviciul web la un moment dat şi care de altfel, era şi instabil. Cu Apache a trebuit să utilizăm modulul Fast CGI[8] care permite serverului de web să încarce *o singură dată* o instanţă a serviciului web (moment în care se încarcă toate resursele necesare – operaţie costisitoare ca timp de execuţie) iar apoi să servească mai mulţi clienţi folosind această instanţă. De asemenea, FCGI permite distribuţia uniformă a încărcării pe instanţele serviciului web existente şi un întreg management al pornirii/opririi instanţelor serviciului web.

Serviciul web TTL pentru platforma WebLicht expune câte un URL pentru fiecare operaţie elementară pentru fiecare limbă în parte (`{lang}` poate lua una din valorile engleză (`en`), română (`ro`) sau franceză (`fr`)):

- segmentare la nivel de frază şi apoi la nivel lexical: `http://ws2.racai.ro/TTL-{lang}-tokenizer`;

[6] Abrevierea din engleză pentru „Tokenizing, Tagging and Lemmatizing free running texts”.
[7] http://httpd.apache.org/
[8] FCGI, http://www.fastcgi.com/

- adnotare cu etichete morfo-sintactice: `http://ws2.racai.ro/TTL-{lang}-postagger`;
- lematizare: `http://ws2.racai.ro/TTL-{lang}-lemmatizer`;
- analiza sintactică de suprafaţă: `http://ws2.racai.ro/TTL-{lang}-chunker`;

În afară de aceste operaţii, orice serviciu web din WebLicht necesită conversia textelor primare (care nu sunt adnotate în vreun fel) la formatul TCF de bază (fişierul XML care este preluat de prima operaţie care, de obicei, este segmentarea la nivel de frază/unitate lexicală). În prezent, WebLicht conţine astfel de servicii web care convertesc texte UTF-8 şi RTF la formatul TCF de bază exemplificat în figura 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.3">
  <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata">
    <tns:source></tns:source>
  </tns:MetaData>
  <tns:TextCorpus xmlns:tns="http://www.dspin.de/data/textcorpus" lang="ro">
    <tns:text>Aceasta este o propoziţie de test.</tns:text>
  </tns:TextCorpus>
</D-Spin>
```

Figura 1: Formatul TCF pentru propoziţia „*Aceasta este o propoziţie de test.*”

Figura 2 reprezintă o sesiune de lucru în WebLicht. Primul pas pe care trebuie să-l facă utilizatorul este să selecteze limba textului pentru care doreşte procesarea. Apoi, alegând orice operaţie din coloana din partea stângă, sistemul va adăuga automat în lanţul de prelucrare toate operaţiile care aduc textul din forma brută în formatul TCF necesar la intrarea operaţiei selectate de utilizator. De exemplu, dacă am fi selectat operaţia de adnotare cu etichete morfo-sintactice, WebLicht ne-ar fi adăugat automat în lanţul de prelucrare operaţiile de conversie la formatul TCF de bază (cel din figura 1) şi segmentare la nivel de frază şi unitate lexicală.

Lanţul de prelucrare final apare în chenarul „Selected Tools:” iar apăsarea butonului „Run” produce rezultatele vizibile în partea de jos a coloanei din partea dreaptă. Acestea se pot descărca de pe site în formatul XML TCF produs de lanţul de prelucrare.

Anterior aminteam de posibilitatea extraordinară de a combina diverse operaţii WebLicht pentru a obţine rezultate care altfel (în absenţa acestei platforme) ar fi fost destul de dificil (sau imposibil în cazul în care aplicaţiile nu erau disponibile) de obţinut. Studiul de caz este combinarea rezultatelor a două adnotatoare cu etichete morfo-sintactice care rulează pe acelaşi text segmentat lexical şi frazal în acelaşi fel pentru ambele programe. Tufiş (1999) discută posibilitatea obţinerii unei adnotări morfo-sintactice mai bune în cazul în care acelaşi program de adnotare morfo-sintactică rulează cu modele de limbă diferite. WebLicht ne permite însă să obţinem o adnotare mai bună (aplicând deci aceleaşi procedee ca în lucrarea citată – calculul matricelor de confuzie) folosind două adnotatoare diferite: în cazul nostru, TTL şi TreeTagger pentru limba engleză. În figura 3 avem rezultatele celor două programe pentru propoziţia „*A very simple test sentence is the test bed for the combined classifiers model.*”

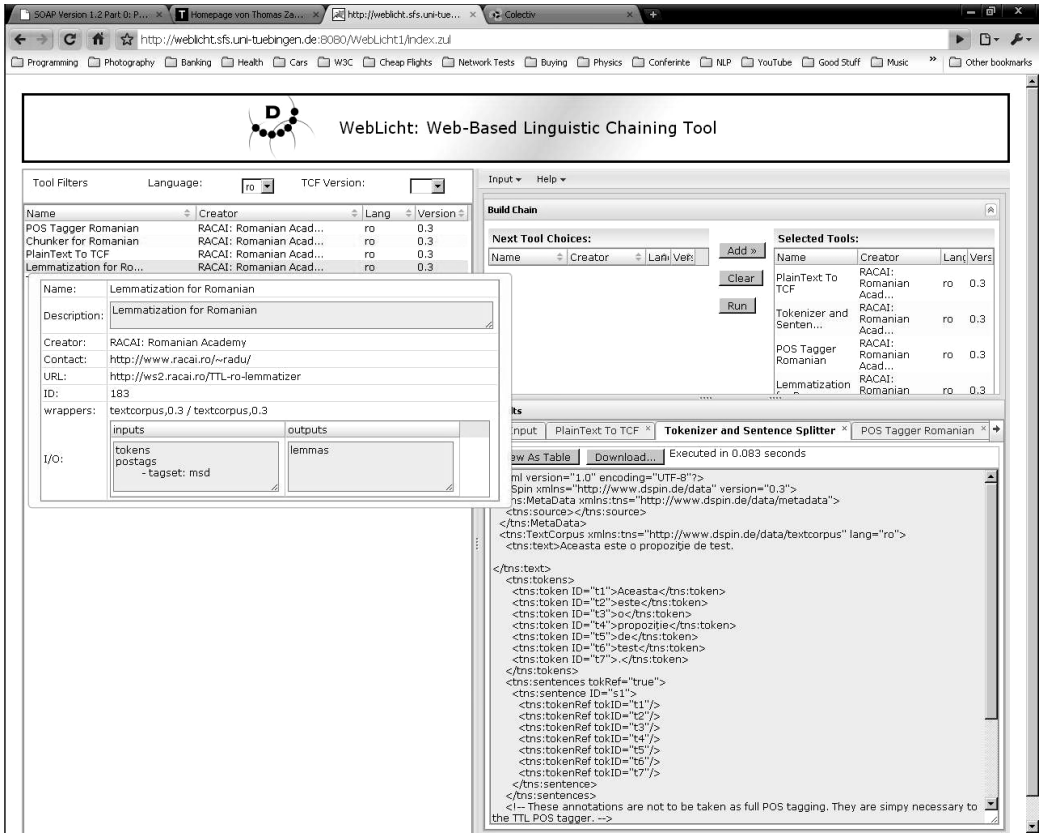


Figura 2: O sesiune de lucru în WebLicht

Pentru a reuși combinarea rezultatelor avem nevoie de corespondența etichetelor folosite (o altă aplicație extrem de utilă pe care WebLicht o face posibilă). TTL folosește mulțimea de etichete Multext-East (MSD) iar TreeTagger-ul folosește mulțimea de etichete Penn Treebank. Știm de exemplu că substantivele la singular se codifică cu „NN" în Penn Treebank și cu „Ncns" în MSD, substantivele la plural cu „NNS" respectiv „Ncnp", adjectivele cu „JJ" respectiv „Afp", ș.a.m.d. În exemplul considerat, ambele programe au găsit adnotarea corectă, dar, în cazul textelor mari, acest lucru este puțin probabil iar cu tehnica din (Tufiș, 1999) (după ce am fixat corespondența etichetelor), putem obține rezultate mai bune.

Tokenization	**TreeTagger**	**TTL**
A	**DT**	**Ti-s**
very	**RB**	**Rsp**
simple	**JJ**	**Afp**
test	**NN**	**Ncns**
sentence	**NN**	**Ncns**
is	**VBZ**	**Vmip3s**
the	**DT**	**Dd**

Tokenization	TreeTagger	TTL
test	**NN**	**Ncns**
bed	**NN**	**Ncns**
for	**IN**	**Sp**
the	**DT**	**Dd**
combined	**JJ**	**Afp**
classifiers	**NNS**	**Ncnp**
model	**NN**	**Ncns**
.	.	**PERIOD**

Figura 3: Rezultatele adnotatoarelor TTL şi TreeTagger pe o propoziţie de test

3. Servicii REST de procesare a textului bazate pe metode statistice de maximizare a entropiei

Pe lângă TTL, ICIA mai dispune de aplicaţii performante de prelucrare primară a textelor care au fost de asemenea adaptate ca servicii web SOAP aşa cum se raportează în (Tufiş et al., 2007). Ca o noutate faţă de aceste servicii web de tip SOAP dezvoltate folosind platforma .Net Framework[9] pentru engleză şi română, noile servicii conforme cu specificaţiile WebLicht (apel de tip REST şi parametri de intrare/ieşire în format TCF) sunt disponibile şi pentru limbile franceză şi germană. Adaptarea la noile limbi a fost posibilă datorită flexibilităţii metodei de antrenare statistică, serviciul REST de adnotare morfo-sintactică[10] folosind modulul de adnotare stratificată METT – Maximum Entropy Tiered Tagging (Ceauşu, 2006), pe principiul maximizării entropiei, similar modelului ME al lui Ratnaparkhi (1988). Ca şi TTL, tagger-ul METT utilizează pentru fiecare limbă un set de descrieri morfo-sintactice compatibil cu specificaţiile Multex-East şi un tagset redus pentru adnotarea stratificată (Tufiş & Dragomirescu, 2004).

Noile servicii web de procesare a textului care implementează METT pentru WebLicht sunt disponibile la adresa:

`http://www.racai.ro/RestWS/Service.svc/ws-{lang}-{webservice}`

unde `{lang}` este limba textului ce urmează a fi procesat, iar `{webservice}` este tipul de procesare ce urmează a fi invocat ca serviciu web. `{lang}` poate lua valorile română (`ro`), engleză (`en`), franceză (`fr`) şi germană (`de`) iar `{webservice}` poate fi înlocuit cu unul din următoarele tipuri de prelucrări: segmentare lexicală (`tokenizer`), segmentare la nivel de frază (`sentsplitter`), adnotare morfo-sintactică (`postagger`) şi lematizare (`lemmatizer`). Trebuie să notăm faptul că este obligatoriu ca în orice interogare HTTP la unul din aceste URL-uri, să specificăm în preambulul interogării

[9] http://www.microsoft.com/NET/

[10] http://www.racai.ro/RestWS/Service.svc/ws-{lang}-postagger unde „{lang}" poate lua valorile „en", „ro", „fr" şi „de".

tipul „application/xml” pentru fişierul TCF pe care îl vom trimite spre prelucrare (în antetul „Content-type”).

Pentru conversia în formatul TCF folosim serviciul web:

`http://www.racai.ro/RestWS/Service.svc/converter-{lang}`

unde `{lang}` este codul ISO de două caractere al limbii textului care urmează a fi procesat. Dacă limba nu este precizată, va fi invocat automat serviciul web pentru recunoaşterea limbii (descris în secţiunea 4).

În figura 4 este prezentat un exemplu de interogare a serviciilor web pentru conversia de la text la formatul TCF, segmentare lexicală şi segmentare la nivel de frază, adnotare morfo-sintactică şi lematizare. În continuare, folosind exemplul din figura 4, descriem parametrii de intrare şi ieşire necesare fiecărui serviciu.

La interogarea serviciului web de conversie a unui text în formatul TCF (http://www.racai.ro/RestWS/Service.svc/converter-ro) rezultatul va fi un document XML similar celui din figura 1. Noul document XML va avea sub elementul „tns:TextCorpus” doar un element „tns:text” cu textul conţinut în interogarea HTTP. În cazul acestui serviciu, parametrul de intrare este de tipul „text/plain” (specificat în „Content-type”) iar cel de ieşire este de tipul „application/xml”.

Pentru următoarea etapă de procesare – segmentarea lexicală – rezultatul serviciului de conversie este folosit ca argument pentru serviciul web http://www.racai.ro/RestWS/Service.svc/ws-ro-tokenizer. Rezultatul interogării HTTP este un document XML care are, pe lângă elementul „tns:text”, şi un alt element „tns:tokens” care conţine atomii lexicali din „tns:text”. Fiecare element „tns:token” conţine indexul din textul iniţial şi lungimea atomului lexical.

În continuare, documentul XML rezultat al serviciului de segmentare lexicală este folosit ca parametru de intrare pentru serviciul de segmentare la nivel de frază (http://www.racai.ro/RestWS/Service.svc/ws-ro-sentsplitter). Acest serviciu adaugă elementul „tns:sentences” în care sunt grupate elemente „tns:sentence” reprezentând frazele formate din atomii lexicali din „tns:tokens”. Serviciul de segmentare frazală necesită prezenţa unui element „tns:tokens” în documentul XML iniţial, deoarece fiecare frază conţine referinţe către atomii lexicali.

În final, rezultatul serviciului de segmentare la nivel de frază este folosit ca parametru de intrare pentru serviciul de adnotare morfo-sintactică stratificată (http://www.racai.ro/RestWS/Service.svc/ws-ro-postagger). Deşi poate folosi ca parametru de intrare un document XML conţinând doar segmentarea lexicală (cu element „tns:tokens”), este recomandat ca documentul XML ce urmează a fi prelucrat să conţină şi segmentarea frazală (cu element „tns:sentences”), informaţia suplimentară contribuind la o viteză crescută a adnotării. Acest serviciu adaugă documentului XML iniţial un nou element „tns:POStags” în care sunt precizate descrierile morfo-sintactice ale atomilor lexicali.

Serviciul web de lematizare (http://www.racai.ro/RestWS/Service.svc/ws-ro-lematizer) necesită rezultatul tokenizării şi al adnotării morfo-sintactice pentru a atribui o lemă

fiecărui atom lexical. Acest serviciu adaugă un element „tns:lemmas” documentului XML inițial.

```
<?xml version="1.0" encoding="utf-8" ?>
- <D-Spin version="3.0" xmlns="http://www.dspin.de/data"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema">
    <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata" />
  - <tns:TextCorpus lang="ro"
      xmlns:tns="http://www.dspin.de/data/textcorpus">
      <tns:text>Acesta este un exemplu.</tns:text>
    - <tns:tokens>
        <tns:token ID="t1" start="0" end="6">Acesta</tns:token>
        <tns:token ID="t2" start="7" end="11">este</tns:token>
        <tns:token ID="t3" start="12" end="14">un</tns:token>
        <tns:token ID="t4" start="15" end="22">exemplu</tns:token>
        <tns:token ID="t5" start="22" end="23">.</tns:token>
      </tns:tokens>
    - <tns:sentences>
      - <tns:sentence ID="s1">
          <tns:tokenRef tokID="t1" />
          <tns:tokenRef tokID="t2" />
          <tns:tokenRef tokID="t3" />
          <tns:tokenRef tokID="t4" />
          <tns:tokenRef tokID="t5" />
        </tns:sentence>
      </tns:sentences>
    - <tns:POStags tagset="MSD">
        <tns:tag tokID="t1">Pd3msr</tns:tag>
        <tns:tag tokID="t2">Vmip3s</tns:tag>
        <tns:tag tokID="t3">Timsr</tns:tag>
        <tns:tag tokID="t4">Ncms-n</tns:tag>
        <tns:tag tokID="t5">PERIOD</tns:tag>
      </tns:POStags>
    - <tns:lemmas>
        <tns:lemma tokID="t1">acesta</tns:lemma>
        <tns:lemma tokID="t2">fi</tns:lemma>
        <tns:lemma tokID="t3">un</tns:lemma>
        <tns:lemma tokID="t4">exemplu</tns:lemma>
        <tns:lemma tokID="t5">.</tns:lemma>
      </tns:lemmas>
    </tns:TextCorpus>
</D-Spin>
```

Figura 4: Exemplu de rezultat al interogării serviciilor web de identificare a limbii, de segmentare lexicală şi frazală, de adnotare morfo-sintactică şi de lematizare.

4. Identificarea limbii

Acest serviciu[11] asigură identificarea automată a limbii unui text scris într-una dintre cele 22 de limbi ale Uniunii Europene. Textul ar trebui să conțină un număr minim de 10-15 cuvinte (în principiu, o propoziție). Serviciul web de identificare a limbii

[11] http://www.racai.ro/RestWS/Service.svc/converter

funcţionează ca un convertor din text în formatul TCF care poate recunoaşte automat limba textului, adăugând un atribut „lang" elementului „tns:TextCorpus".

Identificarea limbii este făcută folosind modele statistice pentru fiecare limbă şi un modul de predicţie. Predicţia este realizată prin calcularea unui scor de similaritate al textului de intrare cu fiecare model de limbă în parte. Modelele de limbă se realizează pe baza ponderii pe care o au prefixele şi sufixele cuvintelor în textele de antrenare disponibile pentru limba respectivă.

În experimentele realizate până acum, am utilizat texte de antrenament (cu mărimi ce au variat între 0,5 şi 1,2 MB) pentru cele 22 limbi oficiale ale Acquis-ului Communautaire (Steinberger et al, 2006). Textele, fiind însă din domeniul juridic şi având o structură mai aparte[12], nu sunt tocmai reprezentative pentru limbile luate în discuţie. Cu toate acestea, am obţinut rezultate excelente folosind pentru prefixe o lungime de trei caractere iar pentru sufixe de patru.

5. Concluzii

De la intrarea lor în funcţiune pe data de 19 februarie 2010, serviciile web ale ICIA din cadrul platformei WebLicht au avut 1389 de accesări pentru toate operaţiile expuse pentru toate limbile. Dintre acestea cele mai multe accesări au fost de test (câte aprox. 40 de bytes de text pe cerere) dar au fost şi cazuri în care s-a cerut procesarea a mai mult de 2KB de text. Suntem astfel siguri că efortul de integrare a aplicaţiilor noastre în standardele promovate de CLARIN va fi de folos atât nouă cât şi comunităţii CLARIN care creşte într-un ritm alert.

Proiectul CLARIN deschide o nouă cale în abordarea cercetărilor în Prelucrarea Automată a Limbajului Natural şi Lingvistică Computaţională prin operaţiunile de standardizare, colectare şi diseminare a unor colecţii impresionante de resurse şi aplicaţii ale acestor domenii care altfel ar fi fost în marea majoritate a cazurilor, inaccesibile. Astfel, noi probleme de cercetare pot să apară sau unele mai vechi îşi pot găsi rezolvarea. În orice caz, adoptarea standardelor CLARIN va asigura accesul neîngrădit al celor neiniţiaţi la tehnologiile limbajului pe de o parte şi îmbunătăţirea considerabilă a şanselor cercetătorilor de a găsi rapid soluţii la problemele lor de cealaltă.

Mulţumiri. Activitatea de cercetare descrisă în această lucrare a fost sprijinită de proiectul european FP7 „CLARIN – Common Language Resources and Technology Infrastructure" (nr. de proiect 212230) finanţat de Comisia Europeană şi de proiectul românesc PC7 „CLARIN–RO: Infrastructură pentru resurse lingvistice interoperabile pentru limba română" (nr. de proiect 16EU/06.04.2009) finanţat de ANCS.

[12] Textele juridice au adesea o structură formată din multe aliniate în care anumiţi termeni se repetă de foarte multe ori afectând acoperirea lingvistică a modelelor de limbă.

Referinţe bibliografice

Ceauşu, A. (2006). Maximum Entropy Tiered Tagging. Janneke Huitink & Sophia Katrenko (editors), *Proceedings of the Eleventh ESSLLI Student Session, ESSLLI 2006*, June 2006, Malaga, Spain, pp. 173—179.

Hinrichs, E., Wittenburg, P., Lemnitzer, L., Geyken, A. (2008). D-SPIN - the German CLARIN initiative. In *CLARIN Newsletter #2*, 2008 (http://www.clarin.eu/files/cnl02-web.pdf).

Kemps-Snider, M., Bel, N., Broeder, D. (2009). Proposal for a CLARIN Service CMDI Components. September, 2009, http://www.clarin.eu/wp2/wg-26/wg-26-documents/cmdi-profile-for-web-services

Kemps-Snider, M., Bel, N. (2009). CLARIN Report on Web Services. March 2009 http://www.clarin.eu/wp2/wg-26

Ratnaparkhi, A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Schmid, H. (2009). The technical details of the D-SPIN Architecture. Internal technical report, 2009 (http://weblicht.sfs.uni-tuebingen.de/englisch/publikationen.shtml).

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4, EAN 9782951740822.

Tufiş, D. (1999). Tiered Tagging and Combined Language Models Classifiers. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Text, Speech and Dialogue (TSD 1999)*, Lecture Notes in Artificial Intelligence 1692, pp. 28—33. Springer Berlin / Heidelberg, January 1999. ISBN 978-3-540-66494-9.

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona, 2004, pp. 39—42.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2007). Servicii Web lingvistice ale ICIA. În Ionuţ Cristian Pistol, Dan Cristea, şi Dan Tufiş, editori, *Lucrările atelierului RESURSE LINGVISTICE ŞI INSTRUMENTE PENTRU PRELUCRAREA LIMBII ROMÂNE*, pp. 61–68, Iaşi, România, 14–15 decembrie 2007. Editura Universităţii „Alexandru Ioan Cuza" Iaşi.

TIPARE HIPONIMICE PENTRU LIMBA ROMÂNĂ

VERGINICA BARBU MITITELU

Institutul de Cercetări pentru Inteligență Artificială, Academia Română

vergi@racai.ro

Rezumat

Hiponimia este o relație lexico-semantică foarte puțin studiată în lingvistica românească, însă ea oferă inginerilor o modalitate foarte eficientă de organizare a materialului lexical util în numeroasele aplicații dezvoltate, care presupun prelucrarea limbajului natural. Prezentăm în această lucrare două modalități de identificare a tiparelor hiponimice din limba română, rezultatele și evaluarea acestora; întrevedem, în final, utilități ale acestor tipare.

1. Introducere

Încă din antichitate au existat preocupări pentru ceea ce numim astăzi relații lexico-semantice. Contextul științific din a doua jumătate a secolului trecut a favorizat un reviriment al interesului pentru acest subiect, datorat conturării tot mai proeminente a unui nou domeniu de cercetare, lingvistica computațională, și a preocupărilor sale de prelucrare și generare a textelor în limbi naturale. Relațiile lexico-semantice intervin, la numeroase niveluri, în înțelegerea și producerea limbajului natural. În plus, ele reprezintă un factor semantic, care, alături de alții (frecvență, factorul stilistico-funcțional, etimologic și cel psihologic), contribuie la organizarea volumului uriaș de cuvinte ce alcătuiesc lexicul unei limbi (Bidu-Vrănceanu, Forăscu, 2005).

Hiponimia este relația paradigmatică de sens ce corespunde celei de incluziune între clase în logică (Lyons, 1977, Cruse, 1986). Cuvântul desemnând clasa mai cuprinzătoare din punct de vedere extensional se numește *hiperonim*, iar cel desemnând clasa cuprinsă este *hiponimul*. În lingvistica teoretică, această relație a fost definită, caracterizată (Cruse, 1986, Murphy, 2003), au fost propuse tipuri (Wierzbicka, 1984, Cruse, 1986, Kleiber, Tamba, 1990), teste de identificare (Cruse, 2004, Nyckees, 1998), a fost analizată prin raportare la relații asemănătoare precum meronimia și instanțierea (Kleiber, Tamba, 1990). În lingvistica computațională, hiponimia este privilegiată în ceea ce privește experimentele de extragere a relațiilor din texte, ei dedicându-i-se cele mai multe cercetări.

În această lucrare prezentăm două modalități de identificare a tiparelor hiponimice din limba română, utile pentru îmbogățirea wordnetului românesc cu noi sinseturi. După o descriere a stadiului actual al identificării hiponimiei în texte (secțiunea 2), descriem două modalități de identificare a tiparelor hiponimice din limba română (subsecțiunile 3.1 și 3.2), comparăm rezultatele celor două abordări (subsecțiunea 3.3), apoi prezentăm rezultatele evaluării acestor tipare pe un corpus de specialitate (secțiunea 4), urmate de prezentarea posibilelor aplicații ale acestui studiu (secțiunea 5) și de concluziile lucrării.

2. Stadiul cercetărilor

Studiile de lingvistică teoretică au adoptat preponderent o perspectivă paradigmatică asupra relaţiilor lexico-semantice. Au fost identificate proprietăţile semantice comune şi enumeraţi termenii (atunci când aceştia erau în număr finit) alcătuind o paradigmă, cu ajutorul analizei semice sau componenţiale. Relaţiile au fost definite, caracterizate, clasificate după diverse criterii. Semanticienii au subliniat, însă, interdependenţa perspectivelor paradigmatică şi sintagmatică în studiul semantic al cuvintelor (Bidu-Vrănceanu, Forăscu, 2005: 30).

Prin natura preocupărilor lor, lingvistica computaţională şi inteligenţa artificială sunt interesate, deopotrivă, şi de planul sintagmatic. Având la dispoziţie cantităţi uriaşe de texte, cercetătorii pot extrage din ele diverse date, printre care şi cuvinte aflate în diverse relaţii. Există două abordări principale ale extragerii relaţiilor din texte: abordarea bazată pe tipare (engl. *pattern-based approach*) şi cea bazată pe gruparea cuvintelor în funcţie de sensul lor (engl. *clustering approach*). În cazul celei dintâi, există presupoziţia că, la nivelul textului, pot fi identificate structuri lexico-sintactice cu grad înalt de specificitatea în instanţierea unei perechi de cuvinte aflate într-o anumită relaţie semantică. Cele mai multe experimente s-au efectuat pentru hiponimie (Hearst, 1992, Alfonseca, Manandhar, 2001, Mann, 2002, Paşca, 2004, Oakes, 2005, Pantel, Pennacchiotti, 2006, Barbu Mititelu, 2008 etc.), meronimie (Berland, Charniak, 1999, Gîrju et al., 2006, Pantel, Pennacchiotti, 2006) şi alte relaţii precum persoană – data naşterii, inventator – invenţie, descoperitor – descoperire (Ravichandran, Hovy, 2002) etc. Abordarea bazată pe gruparea cuvintelor s-a concentrat asupra hiponimiei (Caraballo, 1999, Pantel, Ravichandran, 2004, Pantel, Pennacchiotti, 2006). Lucrarea de faţă se înscrie în primul tip de abordare.

3. Identificarea tiparelor hiponimice din limba română

În această lucrare folosim termenul *tipare hiponimice* pentru a ne referi la acele structuri lexico-sintactice care permit coocurenţa în text a unui hiperonim şi a unuia dintre hiponimele sale, la scurtă distanţă unul de altul, astfel: fie HIPERONIM *tipar hiponimic* HIPONIM, fie HIPONIM *tipar hiponimic* HIPERONIM. Am creat acest termen pentru a desemna ceea ce în literatura internaţională de specialitate este numit *hyponymy/hyponymic pattern*.

Experimentele de identificare a tiparelor hiponimice din limba engleză sunt numeroase (vezi mai sus referinţele). Din cunoştinţele noastre, pentru limba română nu a mai efectuat nimeni asemenea experimente.

Prezentăm în continuare două modalităţi în care am identificat în texte româneşti tipare hiponimice.

3.1 Identificarea semi-automată a tiparelor hiponimice din limba română

Folosind un script Perl, am extras dintr-un corpus românesc (segmentat, lematizat, 881817 unităţi lexicale) propoziţiile care conţin substantive aflate în relaţie de

hiponimie (directă sau indirectă[1]), indiferent de distanţa la care se află unul de celălalt în text. Pentru recunoaşterea acestora, am utilizat wordnetul românesc, RoWN (Tufiş et al. 2008), care avea, la momentul respectiv, 46.269 de sinseturi. Nu am impus nicio restricţie asupra distanţei din arborele hiperonimic din wordnet. Propoziţiile astfel extrase le-am grupat, automat, după asemănarea materialului lexical dintre hiponim şi hiperonim: dacă în *n* propoziţii apar un hiponim şi un hiperonim (oricare ar fi ele) al său şi între ele se află acelaşi grup de cuvinte, cu leme identice, în aceeaşi ordine, cele *n* propoziţii, indiferent de hiponimul şi hiperonimul conţinute, sunt grupate împreună, ca reprezentând exemple ale aceluiaşi tipar hiponimic. Am evidenţiat, în felul acesta, următoarele tipare (în care GN reprezintă un grup nominal, h=hiponim şi H=hiperonim):

- *GN(H) **şi anume** GN(h): Ea nu are nici_o bază de susţinere, numai o **idee**, şi anume o **idee** indestructibilă.*[2]
- *GN(h) **fi un fel de** GN(H): **exprimare** e un_fel_de **pornografie***
- *GN(H) **care avea fi** GN(h)**: oameni** care au fost **participanţi***
- *GN(h) **şi (orice) alt** GN(H)**: bani** şi alte **lucruri** suspecte; **taxele** şi orice alte **venituri** ale bugetului de stat*
- *GN(h) **şi (tot) celălalt** GN(H)**: Legile** şi toate celelalte **acte** normative*
- *GN(H), **mai_ales** GN(h)**: delicvenţilor** de drept comun, mai_ales **gangsterilor***
- *GN(h) **deveni** GN(H): s-ar putea ca anumite **prevederi** să devină **subiect** de dispută*
- *GN(H) **nu ca un** GN(h): se poarte ca un **om**, nu ca un **primar***
- *GN(h) **fi considerat (ca)** GN(H): **televiziunile** sunt considerate ca principale **instrumente** de luptă politică*
- *GN(H) **sine numi** GN(h): Aceşti **oameni** se numeau **capitalişti***
- *GN(H), **inclusiv** GN(h)**:** toţi **oamenii** puterii, inclusiv **miniştrii***
- *GN(h) **fi un** GN(H)**: Turcul** e un **om** sărman, are întotdeauna mustaţă, face bani din piatră seacă şi te întreabă dacă îţi place Turcia*
- *GN(h) **sau alt** GN(H): bănci **sau alte** instituţii de împrumut*
- *GN(H), **adică** GN(h): revin la „acest **subiect** sensibil”, adică la **cazul** Vântu - FNI*
- *GN(H), **ci (şi/doar) un** GN(h)**:** e un **om** cu idei, ci doar un **animal***

Am testat aceste tipare pe un corpus jurnalistic de 900.000 de unităţi lexicale. Am extras automat din corpus acele propoziţii în care apar structurile de mai sus. Manual, le-am selectat doar pe cele care îndeplineau şi criteriul sintactic. Rezultatele sunt prezentate în Tabelul 1. Prin „număr de apariţii” înţelegem numărul total de apariţii ale structurii respective în corpus. Prin „număr de apariţii relevante” desemnăm numărul de situaţii în care în poziţiile celor două grupuri nominale apar cuvinte aflate în relaţie de hiponimie.

[1] Hiponimia directă se stabileşte între două noduri aflate în relaţie, din care unul este nodul-tată, iar celălalt nodul-fiu. Hiponimia indirectă se stabileşte, datorită tranzitivităţii acestei relaţii, între noduri care nu sunt în relaţie unul cu celălalt, decât prin intermediul altui nod.

[2] Exemplul nu trebuie să ne surprindă, întrucât, în WordNet, cuvântul polisemantic *idee* apare în mai multe sinseturi, dintre care două sunt în relaţie de hiponimie: ex.: sinsetul {concept 1, idee1, noţiune2} cu glosa „idee generală care reflectă just realitatea” este hiponim al sinsetului {gând1, idee3} cu glosa „rezultatul procesului de gândire”.

Această relație poate fi deja înregistrată în RoWN, poate fi valabilă, deși, RoWN, incomplet fiind, nu o conține, sau poate fi o hiponimie strict contextuală, care nu necesită înregistrare în rețeaua semantică. Am efectuat manual această etapă de evaluare constrânși de două circumstanțe: incompletitudinea RoWN, pe de o parte, și dorința de a nu pierde din vedere cazurile de hiponimie stabilită strict contextual, care nu ar fi fost recunoscute automat. În plus, întrucât corpusul cu care am lucrat nu este parsat, evaluarea manuală a permis și luarea în considerare a acelor cazuri în care determinanți ai hiponimului sau hiperonimului se intercalează în structură, neafectând caracterul acesteia: de exemplu, „revin la „acest **subiect** sensibil", adică la **cazul** Vântu - FNI", unde hiperonimul *subiect* este urmat de un determinant (adjectivul *sensibil*), iar hiponimul *cazul* este precedat de prepoziția *la*.

Precizia tiparelor am calculat-o ca raport procentual între numărul relevant de ocurențe ale unui tipar și numărul său total de ocurențe. În cazul acestei analize, precizia este egală cu acoperirea (engl. *recall*), deci reprezintă chiar acuratețea acestor tipare.

Tabel 1: Rezultatele testării tiparelor hiponimice românești

Nr. crt.	**Tipar**	**Număr de apariții**	**Număr de apariții relevante**	**Acurateț̦e (%)**
1.	GN și orice alt GN	2	2	100
2.	GN și celălalt GN	4	4	100
3.	GN mai_ales GN	1	1	100
4.	GN fi considerat GN	3	3	100
5.	GN sine numi GN	6	6	100
6.	GN fi un GN	36	36	100
7.	GN sau alt GN	2	2	100
8.	GN ci (și/doar) un GN	3	3	100
9.	GN deveni GN	15	14	93,3
10.	GN și anume GN	11	10	90,1
11.	GN și alt GN	7	6	85,7
12.	GN inclusiv GN	31	23	74,2
13.	GN adică GN	8	5	62,5
14.	GN nu ca un GN	1	0	0

Menționăm că două tipare nu s-au regăsit în corpusul folosit pentru evaluare: *GN fi un fel de GN*, *GN care avea fi GN*.

3.2 Traducerea tiparelor englezești și regăsirea lor în corpusuri românești

În (Barbu Mititelu, 2008) am prezentat un experiment în care am identificat tipare hiponimice englezești. Redăm, în Tabelul 2, tiparele hiponimice englezești cele mai relevante și acuratețea lor (NP înseamnă *noun phrase* „grup nominal").

Tabel 2: Tipare hiponimice englezești și acuratețea lor

Nr. crt.	**Tipar**	**Acurateț̦e (%)**
1.	NP other than NP	100
2.	NP especially NP	100

Nr. crt.	Tipar	Acurateţe (%)
3.	NP principally NP	100
4.	NP usually NP	100
5.	NP such as NP	99,2
6.	NP in particular NP	92,3
7.	NP e(.)g(.)NP	91,4
8.	NP become NP	91
9.	NP another NP	87
10.	NP notably NP	86,8
11.	NP particularly NP	84,6
12.	NP except NP	84,6
13.	NP called NP	81,5
14.	NP like NP	81,3
15.	NP including NP	80,6
16.	NP mainly NP	75
17.	NP mostly NP	70,8
18.	NP i.e. NP	65

Menţionăm că în afară de tiparele incluse în tabel, am mai identificat şi altele, pentru care, însă, nu am putut stabili acurateţea, întrucât ele nu s-au regăsit în corpusul de testare. Este vorba despre tiparele: NP *be another* NP, NP *namely* NP, NP *and other* NP, NP *or other* NP, NP *a form of* NP, NP *or another* NP, NP *and similar* NP, NP *or similar* NP, NP *not least* NP, NP *but not* NP, NP *a kind of* NP, NP *like other* NP, NP *in common with other* NP, NP *and sometimes other* NP, NP *and many other* NP, NP *and in other* NP, NP *or any other* NP, NP *which be* NP, NP *for example* NP, NP *that is* NP, NP *apart from* NP, NP *even* NP, NP *be* NP, NP *for instance* NP, NP *as* NP, NP *either* NP, NP *as well as* NP. Am tradus aceste tipare şi am verificat apariţia lor într-un corpus românesc de 900000 de unităţi lexicale. Metoda de evaluare este identică cu cea descrisă mai sus, pentru tiparele româneşti identificate semiautomat. Rezultatele sunt înscrise în Tabelul 3.

Tabel 3: Tipare hiponimice româneşti obţinute prin traducerea celor englezeşti şi ocurenţa lor în corpus

Nr. crt.	Tipar	Număr de apariţii	Număr de apariţii relevante	Acurateţe (%)
1.	GN de exemplu GN	1	1	100
2.	GN ca de pildă GN	1	1	100
3.	GN de pildă GN	2	2	100
4.	GN cum ar fi GN	7	7	100
5.	GN, mai puţin GN	1	1	100
6.	GN de obicei GN	1	1	100
7.	GN altul/alta/alţii/altele decât GN	2	2	100
8.	GN sau alt/altă/alţi/alte GN	9	9	100
9.	GN sau orice alt/altă/alţi/alte GN	2	2	100
10.	GN şi anume GN	2	2	100
11.	GN numit GN	51	50	98

Nr. crt.	Tipar	Număr de apariţii	Număr de apariţii relevante	Acurateţe (%)
12.	GN deveni GN	114	101	88,6
13.	GN mai ales GN	8	7	87,5
14.	GN şi alt/altă/alţi/alte GN	53	44	83
15.	GN adică GN	30	19	63,3
16.	GN cu excepţia GN	16	10	62,5
17.	GN care fi GN	8	5	62,5
18.	GN în special GN	5	3	60
19.	GN inclusiv GN	53	29	54,7
20.	GN afară de GN	11	5	45,5
21.	GN precum GN	6	2	33,3
22.	GN în afară de GN	9	2	22,2
23.	GN chiar şi GN	29	6	20,7
24.	GN un fel de GN	27	5	18,5
25.	GN alt/altă/alţi/alte GN	8	1	12,5
26.	GN ca GN	700	40	5,7
27.	GN până şi GN	11	0	0
28.	GN dar nu GN	2	0	0

3.3 Comparaţie între tiparele hiponimice româneşti identificate prin cele două metode diferite

Diferenţa cantitativă între tiparele traduse şi cele identificate îşi găseşte trei explicaţii:

- diferenţa cantitativă între cele două wordneturi utilizate: wordnetul englezesc PWN 2.1 conţine 117.597 de sinseturi, în vreme ce versiunea de RoWN folosită de noi conţine 46.269 de sinseturi;
- diferenţa cantitativă între corpusurile utilizate: fragmentele din British National Corpus pe care le-am folosit pentru identificarea tiparelor englezeşti însumează 1863MB, iar cele româneşti 25,6 MB;
- atunci când am evaluat ocurenţele tiparelor traduse, am acceptat ca relevante şi acele situaţii în care se creau hiponimii contextuale, adică hiponimii neînregistrate în wordnet, dar stabilite între cuvinte care, în respectivul context, funcţionează ca hiponime (vezi exemplul 1), precum şi pe acelea în care hiponimia rezultă în urma decodării metonimice a contextului[3] (vezi exemplul 2):

(1) ... să comercializeze chiloţi, târlici, oale şi alte comedii...

(2) Pentru cei care trec prin guvern, parlament şi alte funcţii publice, fenomenul de combatere a sărăciei se încheie cu o inexplicabilă îmbogăţire.

[3] Hearst (1992) notează şi ea existenţa unor hiponimii rezultate din metonimie, dependente de context sau de perspectiva din care sunt privite lucrurile. Având în vedere scopul cu care extrage ea hiponime din text, adică acela de a îmbogăţi automat WordNetul englezesc, problema nu este deloc trivială şi necesită o cântărire a exemplelor înainte de a le introduce în WordNet.

Comparând mulțimea tiparelor românești identificate prin traducerea celor englezești (3.2) cu mulțimea celor identificate semiautomat (3.1), constatăm următoarele:

- intersecția celor două mulțimi, reprezentată de contextele identificate prin ambele metode de lucru, cuprinde: *sau alt*, *și anume*, *sine numi*, *deveni*, *mai_ales*, *și alt*, *adică*, *inclusiv*. Aceasta înseamnă că aproape 29% din tiparele traduse au fost identificate și semiautomat și că 57% din cele regăsite semiautomat au fost identificate și prin traducerea tiparelor englezești;
- aceste tipare comune celor două mulțimi sunt, așa cum era de așteptat, dintre cele cu acuratețe mare; în general, peste 80%, excepție făcând *adică* și *inclusiv*. Calculând o medie a preciziilor finale pentru fiecare dintre acestea, se pare că tiparul *sau alt* are precizie maximă; celelalte au o precizie medie de peste 90%: *numit/sine numi* 99%, *și anume* 95%, *mai_ales* 93,8%, *deveni* 91%; urmează apoi *și alt* cu 84,4%, *inclusiv* cu 64% și *adică* cu aproape 63%. În termeni lingvistici, cu cât precizia este mai mică, cu atât tiparul respectiv manifestă o „polisemie sintactică" mai mare, adică apare în mai multe structuri sintactice, cu valori diferite.
- tiparele comune nu sunt neapărat și cele mai frecvente în limbă, lucru iarăși așteptat. Precizia mare a acestora implică un polisemantism sintactic redus, deci o frecvență scăzută.

Dacă facem o comparație între tiparele englezești cu precizie maximă și cele românești cu precizie maximă, constatăm că echivalenții a trei dintre cele englezești se regăsesc printre tiparele cu precizie maximă din limba română: *other than – altul decât*, *especially – mai_ales*, *usually – de obicei*.

Din numărul total de tipare englezești, 70% au ca echivalent în limba română tipare hiponimice. Invers, din numărul total de tipare românești, 66,(6)% au ca echivalent în limba engleză tipare hiponimice. După cum se observă, procentele sunt foarte apropiate, ceea ce dovedește faptul că limbi diferite tind să permită coocurența în contexte similare a cuvintelor aflate în relație de hiponimie.

4. Evaluarea tiparelor hiponimice românești pe un corpus specializat

Am testat tiparele hiponimice din limba română pe un subcorpus al corpusului OPUS (http://www.let.rug.nl/~tiedemann/OPUS/), adică pe documentele EMEA (European Medicines Agency), cuprinzând 11.914.802 de unități lexicale și semne de punctuație. Acest corpus are două caracteristici care influențează rezultatele experimentului: abundența expresiilor repetate și vocabularul specializat (Tiedemann, 2009). Metoda evaluării este aceeași cu cea descrisă mai sus pentru evaluarea tiparelor identificate prin cele două metode diferite (vezi 3.1. și 3.2). Rezultatele sunt cuprinse în Tabelul 4.

Tabel 4. Testarea tiparelor hiponimice pe un corpus specializat

Tipar	**Acuratețe (%)**
GN *chiar și* GN	100
GN *de obicei* GN	100

Tipar	Acurateţe (%)
GN, *ci (şi/doar)* GN	100
GN *în special* GN	96.88
GN *precum* GN	94.83
GN *cum ar fi* GN	93.75
GN *(în) afară de* GN	92.11
GN *şi (orice) alt* GN	90.1
GN *fi un* GN	87.98
GN *sau alt* GN	86.96
GN *mai ales* GN	85.71
GN *alt decât* GN	85.71
GN *sine numi* GN	84
GN *inclusiv* GN	83.51
GN *de exemplu* GN	79.57
GN *fi considerat* GN	79.17
GN *care fi* GN	74.12
GN, *adică* GN	66.66
GN *cu excepţia* GN	54.55
GN *şi (tot) celălalt* GN	54.29

Comparând aceste rezultate cu cele obţinute pe baza corpusurilor jurnalistice, remarcăm că, în majoritatea cazurilor, acurateţea scade. Sunt şi câteva tipare care dovedesc o acurateţe mai mare în cazul textelor specifice unui domeniu: GN, *adică* GN, GN *care fi* GN, GN *în special* GN, GN *inclusiv* GN, GN *precum* GN, GN *(în) afară de* GN. Două dintre tipare au acurateţe maximă în ambele tipuri de texte: jurnalistic şi specializat: GN *de obicei* GN şi GN, *ci (şi/doar)* GN.

5. Aplicaţii ale tiparelor hiponimice

Wordnetul românesc este o resursă lingvistică extrem de utilă aplicaţiilor din lingvistica computaţională. Calitatea şi cantitatea sinseturilor conţinute contribuie la calitatea aplicaţiilor care folosesc wordnetul. În aceste condiţii, îmbogăţirea lui este una dintre priorităţile noastre. Dezvoltat manual până acum, i se pot adăuga sinseturi identificate automat cu ajutorul tiparelor hiponimice prezentate mai sus, după ce un specialist lingvist le adaugă o definiţie şi le verifică completitudinea.

Tiparele hiponimice pot fi utilizate pentru îmbogăţirea RoWN, pe de o parte, atât cu hiponime, cât şi cu instanţe, iar pe de altă parte, atât cu cuvinte din vocabularul general, cât şi cu termeni din diverse domenii.

În cercetarea prezentată aici, nu am distins între hiponimie şi instanţiere, întrucât ele erau tratate identic în versiunea de wordnet folosită. Acestea sunt, însă, relaţii semantice distincte, iar ultima versiune de wordnet englezesc, Princeton WordNet 3.0, le marchează diferit. Rămâne de cercetat în ce măsură tiparele identificate astfel îşi modifică precizia dacă facem această distincţie şi cât de precise sunt în identificarea instanţelor. Includerea instanţelor într-o resursă lingvistică precum wordnetul îşi dovedeşte cu prisosinţă utilitatea în sarcini de identificare a entităţilor numite (engl. *named entities recognition*).

În ceea ce priveşte cuvintele din limba comună, marea majoritate a acestora au fost deja incluse în RoWN. Interesul îl poate reprezenta îmbogăţirea sa cu termeni din diverse domenii. Experimentul de testare a tiparelor hiponimice pe un corpus specializat aduce date relevante în această privinţă.

Din perspectiva semanticii lexicale, experimente ca cel prezentat aici pot aduce completări, exemplificări şi rafinări analizei preponderent paradigmatice a hiponimiei.

6. Concluzii

Lucrarea de faţă se înscrie pe linia cercetărilor de extragere a relaţiilor semantice din corpusuri cu ajutorul tiparelor lexico-sintactice. Am prezentat două modalităţi de identificare a tiparelor hiponimice din limba română. Una dintre metode este semiautomată, cealaltă presupune traducerea unor tipare hiponimice englezeşti şi verificarea echivalentelor acestora în texte româneşti. Am calculat acurateţea tiparelor atât într-un corpus jurnalistic, cât şi într-unul dintr-un domeniu.

Experimentul nu este deloc lipsit de utilitate: tiparele hiponimice pot fi folosite pentru identificarea în corpusuri a hiponimelor de introdus în wordnetul românesc, şi chiar şi a instanţelor.

Mulţumiri. Autoarea este recunoscătoare Ministerului Educaţiei, Cercetării, Tineretului şi Sportului, finanţator al proiectului SIR-RESDEC, în cadrul căruia s-a derulat cercetarea prezentată în acest articol.

Referinţe bibliografice

Alfonseca, E., Manandhar, S. (2001). Improving an Ontology Refinement Method with Hyponymy Patterns. *Third International Conference on Language Resources and Evaluation*, Las Palmas, 235-239.

Barbu Mititelu, V. (2008). Hyponymy Patterns. Semi-automatic Extraction, Evaluation and Inter-lingual Comparison. *Text, Speech and Dialogue*, Springer (P. Sojka, A. Horak, I. Kopecek, P. Karel, eds.), 37-44.

Berland, M., Charniak, E. (1999). Finding parts in very large corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 57-64.

Bidu-Vrănceanu, A., Forăscu, N. (2005) *Limba română contemporană. Lexicul*, Bucureşti, Humanitas Educaţional.

Caraballo, S. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text. *Proceedings of the 37th Annual Meeting of the Association for Computational Liguistics*, 120-126.

Cruse, D. A. (1986). *Lexical Semantics*, Cambridge, CUP.

Cruse, D. A. (2004). *Meaning in Language. An Introduction to Semantics and Pragmatics*, ediţia a doua, Oxford, OUP.

Gîrju, R., Badulescu, A., Moldovan, D. (2006). Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, 32 (1): 82-135.

Hearst, M. A. (1992). Automated Acquisition of Hyponyms from Large text Corpora. *Proceedings in the Fourteenth International Conference on Computational Linguistics*, Nantes.

Kleiber, G., Tamba, I. (1990). L'hyponymie revisité: inclusion et hiérarchie. *Langages 98*: L'hyponymie et l'hyperonymie, Larousse.

Lyons, J. (1977). *Semantics*, vol. 1, Cambridge University Press.

Mann, G. S. (2002). Fine-Grained Proper Noun Ontologies for Question Answering. *COLING-02 on SEMANET: building and using semantic networks*, 1-7.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon*, Cambridge, CUP.

Nyckees, V. (1998). *La sémantique*, Paris, Belin.

Oakes, M. P. (2005). Using Hearst's Rules for the Automatic Acquisition of Hyponyms for Mining a Pharmaceutical Corpus. *Proceedings of the Workshop Text Mining Research, Practice and Opportunities*, Borovets, 63-67.

Pantel, P, Ravichandran, D. (2004). Automatically labeling semantic classes. *Proceedings of HLT/NAACL-04*, Boston, 321-328.

Pantel, P., Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, Sydney, 113-120.

Paşca, M. (2004). Acquisition of Categorized Named Entities for Web Search. *Proceedings of CIKM'04*, Washington.

Ravichandran, D., Hovy, E. (2002). Learning surface text patterns for a question answering system. *Proceedings of ACL-2002*, Philadelphia, 41-47.

Tiedemann, J. (2009). News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing: Selected Papers from RANLP 2007*, John Benjamins, 237-248.

Tufiş, D., Ion, R., Bozianu, L, Ceauşu, A., Ştefănescu, D. (2008). Romanian Wordnet: Current State, New Applications and Prospects. *Proceedings of 4th Global WordNet Conference, GWC-2008*, Szeged (Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen, eds.), 441-452.

Wierzbicka, A. (1984). Apples are not a „kind of fruit". *American Ethnologist 11*, 313-28.

MECANISMELE GENERATIVE ALE MORFOLOGIEI DERIVAȚIONALE

PETIC MIRCEA

Institutul de Matematică și Informatică, Academia de Științe a Moldovei

mirsha@math.md

Rezumat

În articol sunt studiate problemele elaborării unui generator de derivate. Punctul de pornire este un lexicon de cuvinte derivate, care conține nu doar reprezentarea grafică a derivatelor, ci și a morfemelor lor constituente. Aceasta a permis studierea și formularea unor reguli care ar genera cuvinte derivate conform unor restricții la nivel de reprezentare grafică.

Cuvinte-cheie: generare automată a derivatelor, dicționar de derivate, prefix, sufix.

1. Introducere

Dicționarele moderne se confruntă cu unele deficiențe, devenind astfel obiect de cercetare pentru lexicografi. Având în vedere că dicționarele sunt permanent completate cu intrări noi, grație dezvoltării limbii, sarcina elaborării unui vocabular complet rămâne una practic imposibilă. Prin urmare, completarea automată și/sau semiautomată a resurselor lingvistice cu cuvinte generate automat în baza celor deja existente prin mijloace exclusiv interne, în particular cu ajutorul derivatelor cu prefixe și sufixe, reprezintă o sursă semnificativă de îmbogățire a vocabularului.

Pentru automatizarea procesului de derivare este necesar:

- să stabilim regulile, care pot fi aplicate cuvintelor de bază pentru obținerea unor cuvinte derivate noi;
- să stabilim condițiile, în care pot fi aplicate aceste reguli;
- să elaborăm și să aplicăm un mecanism de validare a acestor reguli, având în vedere că restricțiile menționate ulterior nu garantează complet corectitudinea cuvintelor obținute.

Pentru soluționarea problemelor sus menționate este necesar de efectuat un studiu preliminar al procesului de derivare. În calitate de suport lexicografic în acest scop au fost utilizate câteva surse și anume varianta electronică a dicționarului de derivate (Constantinescu, 2008), www.dexonline.ro și RRTLN[1] (*Resurse Reutilizabile ale Tehnologiei Limbajului Natural*).

Scopul acestui articol este de a studia particularităților cuvintelor derivate și de a stabili unele mecanisme generative ale morfologiei derivaționale.

În acest scop, inițial, este descrisă varianta electronică a dicționarului de derivate, evidențiindu-se astfel o caracterizare a dicționarului prin prisma unor date statistice în

[1] Lexiconul se conține pe site-ul http://imi201.math.md/elrr/

ce privește derivatele și constituentele lor. În continuare, este prezentat cazul particular de derivare semianalizabilă, care suscită mai multe semne de întrebare în ce privește modul de formare al lor. Ulterior, s-a studiat situația derivării regulate, care presupune, la rândul său, generarea derivatelor prin schimbarea genului cuvântului, generarea diminutivelor și a augmentativelor. Un compartiment aparte îl constituie procesul de proiectare a derivării cu prefixe cu o ulterioară derivare cu sufixe. În final sunt prezentate unele particularități ale generării automate a derivatelor cu prefixele *im-*/*in-*.

2. Lexiconul cu derivate

Lexiconul reprezintă varianta electronică a dicționarului de derivate, elaborat de S. Constantinescu (Constantinescu, 2008) și conține doar reprezentarea grafică a derivatelor și morfemelor constituente, fără nici o informație despre partea de vorbire a derivatelor și temelor lor. Pentru o procesare mai simplă a intrărilor lexiconului, a fost elaborată o expresie regulată care reprezintă următoarea structură a derivatelor:

```
derivat = (+morfem)*.morfem(–morfem)*
```

unde `+morfem` reprezintă un prefix, `.morfem` este tema și `–morfem` este un sufix. Un exemplu de intrări în dicționar este:

```
antistatal=+anti.stat–al
reprogramabil=+re.programa–bil
```

Au fost elaborați algoritmi și apoi dezvoltate programele (Petic, 2010) care stabilesc caracteristicile statistice ale lexiconului (Tabel 1).

Tabel 1: Caracteristicile statistice ale lexiconului

Caracteristica	**Numărul**
Derivate	15300
Rădăcini	6800
Prefixe	42
Sufixe	433

În momentul procesării lexiconului, prin extragerea grupurilor de derivate cu aceleași afixe, s-a constat că există un grup restrâns de prefixe și sufixe care formează marea majoritate a derivatelor (Petic, 2010).

Astfel, 12 prefixe din cele 42 formează 88,2% din toate derivatele cu prefixe, înregistrate în acest lexicon. Derivatele care sunt formate din următoarele prefixe sunt cele mai numeroase: *ne-*, *re-*, *în-*, *des-*, *pre-*, *anti-*, *auto-*, *sub-*, *dez-*, *supra-*, *de-*, și *îm-*[2].

Pe de altă parte, din 433 de sufixe înregistrate în lexicon 52 formează 87,7% din numărul tuturor derivatelor cu sufixe. Cele mai numeroase derivate s-au dovedit a fi, în ordine descrescătoare, următoarele sufixe: *-re*, *-tor*, *-toare*, *-eală*, *-ie*, *-ătoare*, *-iza*, *-oasă*, *-ar*, *-ător*, *-ească*, *-os*, *-aș*, *-esc*, *-tură*, *-iță*, *-ist*, *-uță*, *-el*, *-i*, *-ui*, *-ătură*, *-ește*, *-ism*, *-a*, *-ărie*, *-ică*, *-ime*, *-itate*, *-ioară*, *-işor*, *-işoară*, *-ic*, *-uleț*, *-că*, *-ean*, *-iș*, *-easă*, *-bil*, *-uț*, *-at*, *-oaică*, *-uşor*, *-an*, *-oi*, *-uliț*, *-iu*, *-enie*, *-istă*, *-al*, și *-ea*.

[2] Derivatele sunt enumerate în ordine descendentă.

E de menționat faptul că nu de la toate temele se formează un număr proporțional de derivate (Petic, 2010). Astfel, sunt cuvinte la care se înregistrează un număr maxim de derivate și anume: **bun** (32 de derivate), **alb** (25 de derivate), **șarpe** (22 de derivate), **roată** (22 de derivate), **om** (20 de derivate). În lexicon, un număr foarte mare de cuvinte (3657) apar cu un singur derivat. O ilustrație a acestei situații este prezentată în Fig.1.

3. Particularitățile derivatelor semianalizabile

După Iorgu Iordan (Iordan, 1970), derivatele în limba română pot fi grupate în: analizabile, semianalizabile și neanalizabile.

În derivatele *analizabile* se recunoaște atât prefixul, cât și cuvântul de bază. În derivatele *semianalizabile* se recunoaște numai prefixul, prin opoziție cu alte derivate sau cuvinte compuse cu o temă comună, inexistentă ca și cuvânt independent (de exemplu: *deschis – închis*). Unele derivate au devenit cu timpul *neanalizabile*, fie din cauza evoluției fonetice, morfologice sau semantice, fie din cauza dispariției cuvântului de bază la care se făcea raportarea.

Depistarea derivatelor semianalizabile pare a fi o problemă mai complicată decât cea a celor analizabile, deoarece temele care au stat la baza derivărilor lor nu mai sunt cunoscute. Astfel, pentru a studia acest proces, este necesar de a avea la dispoziție cuvintele derivate semianalizabile marcate corespunzător. Prin urmare, este util lexiconul, care conține cuvintele derivate împreună cu structura lor morfemică (Constantinescu, 2008). Analizând conținutul lexiconului menționat, s-a constat că el include astfel de derivate, stabilindu-se, că, în mare parte, derivatele semianalizabile se referă la prefixele *des-/în-*. S-a realizat extragerea automată a acestor cuvinte, identificându-se 57 de derivate de acest tip. Așadar, unele merită atenție pentru a pătrunde în esența mecanismului de derivare.

O caracteristică pentru derivatele semianalizabile este utilizarea derivării regresive, adică înlăturarea unui afix și nu adăugarea acestuia, un exemplu elocvent fiind cuvântul *crucișătură* care s-a format prin derivarea regresivă de la verbul *încrucișa* și afixarea cu sufixul *–ătură*. Corectitudinea procesului de formare a acestui cuvânt este confirmată prin intrarea în dicționarul electronic de derivate:

```
crucișătură=.(în)crucișa-ătură
```

Deci, pentru acest caz se observă că intrarea din lexicon descrie modul de formare a derivatului. Verbul *încrucișa* a derivat regresiv, prin înlăturarea prefixului *în-*, ca apoi să fie adăugat sufixul *–ătură*. Intrarea din lexicon a marcat derivarea regresivă prin prefixul luat între paranteze rotunde. În resursa *www.dexonline.ro,* aceeași informație este prezentată după cum urmează:

```
[în]crucișa + suf. -ătură
```

ceea ce reprezintă o marcare similară, cu cea de mai sus, a modului de formare a cuvântului derivat *crucișătură*. Dincolo de acest cuvânt, mai sunt și alte derivate de acest tip, de exemplu, *cingătoare*, *chiondoreală*, *fundătură*, *greunatic*, *lăuntric*, *notătoare*, etc.

Un alt exemplu de derivate semianalizabile sunt derivatele cu prefixele *des-* (*dez-*). Din exemplele ulterioare, se desprind următoarele cuvinte prefixate semianalizabil:

despăduri=+des.(îm)păduri

dezvălui=+dez.(în)vălui

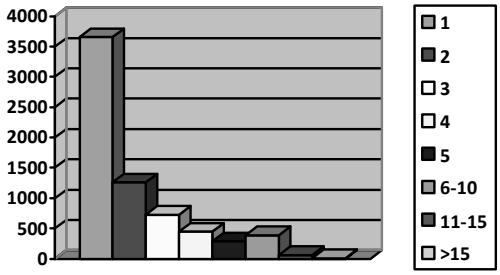


Figura 1: Datele statistice cu referire la numărul derivatelor cu temele din lexicon.

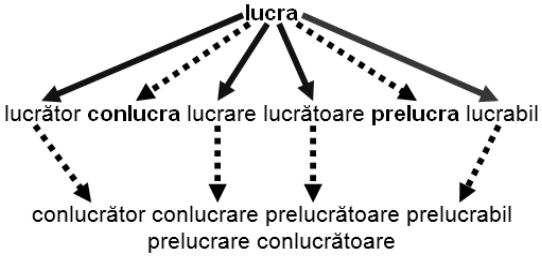


Figura 2: Proiectarea derivatelor în cazul cuvântului *lucra* din limba română.

Astfel se poate afirma că este vorba de un schimb de prefixe, menționându-se, în aceeași ordine de idei, și faptul că nu sunt așa teme precum *păduri*, și *văl*ui. Un moment interesant este că în dicționarul morfologic s-au găsit doar verbe care reprezintă derivate semianalizabile, chiar dacă, în afară de *despăduri*, mai este și substantivul *despădurire*, care nu este găsit ca fiind semianalizabil. În cazul derivatelor resursa *www.dexonline.ro* indică cuvântul de la care a fost derivat, prin marcare cu simbolul *V.* care ar însemna „vezi cuvântul".

4. Proiectarea derivării cu prefixe pe ulterioara derivare cu sufixe

Proiectarea derivatelor reprezintă o metodă de formare a cuvintelor prefixate de la derivate sufixate de la aceeași rădăcină. Potrivit cercetătorilor spanioli (Santana et al., 2004), verbul spaniol *amortizar* poate fi derivat cu prefixul *des-* în *desamortizar*. Totodată *amortizar* poate fi derivat cu sufixele *–cion* și *–able*. Astfel, derivatul cu prefixul *des-* poate deriva cu sufixele *–cion* și *–able*. Apare ipoteza potrivit căreia derivatele cu prefixe pot moșteni/proiecta derivatele cu sufixe ale temei de la care a fost realizată derivarea cu prefixe. Deci, mecanismul de generare a derivatelor ar presupune următoarele: dacă R este rădăcina cuvintelor, S_i sufixele posibile pentru rădăcina R, adică $R \rightarrow RS_i$, și P prefixul corespunzător rădăcinii R, adică $R \rightarrow PR$, atunci există sufixe S_i pentru care $R \rightarrow PRS_i$

Prin urmare, dacă luăm drept exemplu tema *capitula*, atestăm conform *www.dexonline.ro* următoarele derivate cu sufixe: *capitulant*, *capitulantă*, *capitulard*, *capitulare*, *capitulație*. Totodată, pentru derivatul cu prefixul *re-*, *recapitula* are următoarele derivate cu sufixe: *recapitulare*, *recapitulație*. Astfel, sunt două derivate care realizează o proiectare a derivării cu sufixe. Încercând proiectarea derivatului *capitulant* → *recapitulant,* se obține un cuvânt inexistent, chiar și pentru resursele electronice ale motorului de căutare *www.google.com*.

Pornind de la lexiconul existent cu structura morfemică (Constantinescu, 2008), cu ajutorul unor programe, s-au extras acele grupuri de derivate la care se atestă proiectarea, înregistrându-se 363 de teme de la care este posibilă proiectarea derivatelor.

Marea majoritate a acestora constă doar dintr-un derivat sufixat şi unul prefixat. Totuşi, se poate afirma că metoda este utilă în generarea derivatelor pentru limba română. Un exemplu de astfel de derivare este ilustrat în Figura 2.

5. Studierea diminutivelor şi a augmentativelor, derivarea prin schimbarea genului

Potrivit cercetătorilor sârbi (Duško&Krstev, 2005), cu ajutorul cuvintelor derivate din limba sârbă, au fost generate noi leme cu sensuri previzibile. Prin sensuri previzibile se înțelege derivarea prin amplificarea sensului, adică generarea diminutivelor (*profesorčić*) şi a augmentativelor (*profesorčina*), schimbarea genului cuvântului (pentru cel masculin *professor* → cel feminin *professorka*),

Procesarea diminutivelor

Sufixe diminutivale ar fi următoarele: *-aș* (copilaș), *-uc* (sătuc), *-el* (bătrânel), *-iță* (fetiță), *-uță* (caruță), *-ică* (floricică), *-uleț* (ursuleț), *-iș* (podiș), *-uț* (căluț).

Tabel 2: Numărul de derivate diminutive pentru sufixe concrete

Sufixul	**Numărul de derivate**
-aș	327
-iță	249
-el	221
-uță	208
-ică	139
-uleț	104
-iș	101
-uț	88
-uc	7

Tabel 3: Numărul de derivate augmentative pentru sufixe concrete

Sufixul	**Numărul de derivate**
-andru	4
-an	74
-oi	74
-oaie	26

Potrivit aceleaşi surse (Constantinescu, 2008), au fost stabilite numărul de derivate cu fiecare sufix diminutival (Tabel 2.). Dincolo de aceasta, toate aceste sufixe înregistrează unele alternanțe vocalice şi/sau consonantice în cazul procesului de derivare, de exemplu, *sa**ă**t - sătuc, ca**ă**r - căruță, cal - căluț* unde alternanța este *a->ă; fa**ț**ă – fe**ț**iță: a ->e; flo**oa**re – flo**o**ricică: oa ->o; fra**ț**e- fră**ț**ior: at -> ăț* etc.

Cea mai numeroasă clasă formată în urma procesului de derivare cu aceste sufixe este clasa substantivelor, menționându-se aici următoarele sufixe *–uc*, *-el*, *-aș*, *-iș*, *-uț*, *-uță*, *-iță*, *-uleț* şi *-ică*. Mai puțin numeroase sunt adjectivele, în cazul sufixelor: *-uc*, *-el*, *-iș*, *-uț* şi *–ică.* Un caz foarte rar s-a atestat cu sufixul *-iș*, care a format de la verbul *zbura* adverbul *zburiș*.

Procesarea augmentativelor

Sufixe augmentative sunt următoarele: *-andru* (copilandru); *-an* (băietan); *-oi/oaie* (căsoi, căsoaie). Potrivit lexiconului (Constantinescu, 2008), au fost stabilite numărul de derivate cu fiecare sufix augmentativ (Tabel 3.). Toate aceste sufixe pot fi atașate substantivelor pentru a obține substantive. Sufixele *–an*, *-oi*, *-oaie* pot fi atașate

adjectivelor pentru a obține cuvinte augmentative. De menționat că în procesul de derivare cu sufixele respective se atestă unele alternanțe vocalice și/sau consonantice, cum ar fi de exemplu, *c**a**să - c**ă**soi/c**ă**soaie,* unde alternanța este *a->ă, băie**t** – băie**ț**andru t->ț,* etc.

Schimbarea genului cuvântului

În cazul limbii române, schimbarea genului cuvântului poate fi realizată prin trecerea de la unele sufixe la altele corespunzătoare, de exemplu, *-tor* ↔ *-toare*, *-esc* ↔ *-ească*, etc. Astfel, se observă că la schimbarea genului se apelează la procesul de sufixare, și nu la cel de prefixare.

Lexiconul menționat anterior conține derivate care sunt sufixate cu *–tor*, cu *–toare* și atât cu *–tor* cât și cu *–toare*. Astfel, potrivit informației din Figura 3, există 148 de cuvinte (substantive și/sau adjective) de forma ω′=ω*tor*, care ar putea deriva în cuvinte de forma ω″=ω*toare*. Similar, sunt 42 de cuvinte (substantive și/sau adjective) de forma β′=β*toare* care e posibil să deriveze în β″=β*toare*. Totodată, aceste 190 de derivate generate automat urmează să fie validate. Pentru început cuvintele au fost verificate la prezența lor în RRTLN. Din toate cuvintele generate 122 au fost prezente pe când restul cuvintelor au fost verificate în documentele electronice din Internet, din 68 de derivate, fiind validate 49. Prin urmare 95% din cuvintele generate au fost valide.

Aceeași situație este și cu perechea de sufixe *–esc* și *–ească*. Potrivit aceluiași lexicon (Constantinescu, 2008), se conțin 274 de derivate cu sufixul *–esc*, și 249 cu *–ească*. De menționat că sunt 229 de derivate care sunt sufixate atât cu *-esc* cât și cu *–ească*. Devine firească ipoteza potrivit căreia cuvintele (substantive și/sau adjective) de forma ω′=ω*esc*, pot deriva în cuvinte de forma ω″=ω*ească*. Similar, cuvintele (substantive și/sau adjective) de forma β′=β*ească* pot deriva în cuvinte de forma β″=β*esc*. Generând automat acele derivate care lipsesc în cazul genului și verificând automat prin existența cuvintelor în documentele electronice, s-a constatat că cu ajutorul RRTLN au fost validate 43 de cuvinte din 65 generate. Din celelalte 22 de derivate s-a reușit de a valida încă 12 derivate cu ajutorul unei aplicații Web, bazate pe posibilitățile motorului de căutare Google. În general s-au obținut 84% de cuvinte valide.

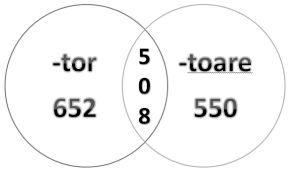


Figura 3: Numărul derivatelor cu sufixele *–tor* și *-toare*

6. Unele aspecte în generarea derivatelor cu in-/im-

Există mai multe clase de derivate cu prefixele *in-/im-*. În cele ce urmează se va descrie cazul în care se formează derivate cu aceste prefixe cu sens negativ. În limba română sunt mai multe prefixe care oferă derivatului sensul de negație, precum ar fi: *a-*, *i-*, *ne-* și *im-/in-*.

Derivatele cu *im-/in-*, de obicei, sunt adjective, rareori substantive sau verbe. Cele mai numeroase derivate cu prefixul *in-/im-* sunt raportate la adjective formate cu sufixul –

bil, de exemplu, *incurabil*, *inestimabil*, etc. Aşadar, fiind date adjectivele $\omega'=\omega bil$ se formează derivatele $\omega''=\beta\omega bil$, unde $\beta\in$ {*in-*, *im-*}.

O altă clasă bine reprezentată este cea a formațiilor raportabile la adjective derivate cu sufixul *–ent* şi *–ant*: *inaderent*, *incoerent*, *independent*, etc. (Iorgu, 1970). La fel, fiind date adjectivele $\omega'=\omega\gamma$, se formează derivatele de forma $\omega''=\beta\omega\gamma$, unde $\beta\in$ {*in-*, *im-*} şi $\gamma\in$ {*-ent*, *-ant*}. În ambele cazuri alegerea lui β depinde de prima literă a adjectivului ω, şi anume în cazul în care această literă este b sau p atunci β=*im-*, în alte cazuri e β=*in-*.

Alte clase de derivate cu *in-/im-* sunt neînsemnate. De menționat că există derivate în *im-/in-* de la teme deja derivate cu prefixe.

Potrivit resursei *www.dexonline.ro*, în limba română sunt circa 4946 de cuvinte care încep cu combinația de litere *in* şi 1249 - cu combinația de litere *im*. Din acestea care se termina cu *ant* sunt corespunzător 38 şi 13, cu *ent* 61 şi 12, şi respectiv *bil* 220 şi 43, în total fiind 387 de cuvinte. Aşa precum unele cuvinte, cum ar fi cele ce se termina cu *bil*, pot aparține mai multor părți de vorbire, s-a reuşit filtrarea şi obținerea unui număr de doar 293 de adjective. Mai mult decât atât, adjectivele cu sufixul *–bil*, mai formează şi derivate cu prefixul *ne-*, care, de altfel, oferă acelaşi sens negativ. În această ordine de idei, este utilă verificarea derivatelor formate cu ajutorul motorului de căutare, de exemplu www.google.com.

Cazul derivării cu prefixul *in-/im-* de la cuvinte care se termină în *–ant* sau *–ent* este şi mai problematic, prin faptul că sunt multe cuvinte care nu vor forma derivate în *in-/im-* pe motiv că sunt şi substantive, nu doar adjective.

Lexiconul de derivate (Constantinescu, 2008) conține doar un singur derivat cu prefixele *im-/in-*, şi anume *impermeabilizare*. Astfel, este posibil de a genera derivate cu prefixele corespunzătoare. În mod automat, s-a constat că lexiconul (Constantinescu, 2008) conține 62 de derivate cu *–bil*, 1 derivat cu *–ent* şi 37 cu *–ant*.

Examinând cuvintele din lexicon şi concatenându-le prefixul respectiv celor care se încadrează în categoriile stabilite, fără alternanțe vocalice sau consonantice, s-a construit un algoritm de derivare cu prefixul *im-/in-*, în baza căruia a fost elaborat un modul care generează cuvinte noi. Drept rezultat s-au obținut 100 de derivate cu prefixele *in-/im-*.

Verificând inițial cuvintele generate la prezența lor în RRTLN, s-a constatat prezența a 7 cuvinte. Celelalte 93 de cuvinte au fost verificate în documentele electronice din Internet cu ajutorul motorului de căutare *www.google.com.* Ca rezultat, doar 14 derivate din toate cele generate au fost regăsite în documentele electronice. Patru cuvinte s-au regăsit o singură data, şi anume: *inofertant, imprelucrabil, inrezolvabil, intrasabil*. Datele pentru restul derivatelor sunt prezentate în Tabel 4.

Tabel 4: Numărul de apariții a derivatelor generate în documentele electronice Internet

Derivatele	**Numărul de apariții**
Invindecabil	7
indefrişabil	8
Indifuzabil	8
Infiltrabil	8

Derivatele	Numărul de apariţii
indeşirabil	56
Insubstituibil	77
Imprecizabil	94
Injustificabil	181
Injucabil	353
Incadrabil	4710

În cazul derivatelor generate cu prefixul *in/im* cu un număr mic de apariţii printre documentele electronice, este clară poziţia potrivit căreia nu a fost aplicat prefixul corect cu sens negativ, aceasta fiind situaţia derivatelor *inrezolvabil* (1) – *nerezolvabil* (1280), *insubstituibil* (77) – *nesubstituibil* (222), *injucabil* (353) - *nejucabil* (3050). Totodată, mai sunt derivate cu un număr mare de apariţii şi care au variante cu alt prefix şi cu un număr mai mic de apariţie a acelui prefix, aici fiind depistate următoarele derivate: *inabordabil* (2810) - *neabordabil* (699), *inacceptabil* (67900) - *neacceptabil* (7140), *incalculabil* (24000) - *necalculabil* (469)

7. Concluzii

Generarea derivatelor nu este o problemă deloc banală, deoarece procesul de derivare a cuvintelor nu prezintă un mecanism regulat. Soluţia potrivit căreia se pot stoca toate derivatele într-un dicţionar este una rezonabilă, or, aceste derivate înregistrate corect oricum nu vor acoperi întreaga diversitate a limbii, care presupune un mecanism în continuă dezvoltare. Pe de altă parte, abordarea potrivit căreia se poate genera derivate conform unor reguli de constrângere a grupurilor de derivate, este un mecanism de supragenerare, în care etapa de validare exclude foarte multe cuvinte formate greşit. Doar stabilirea unor reguli bine puse la punct ar putea ridica nivelul de generare a cuvintelor corecte.

Referinţe bibliografice

Constantinescu, S. (2008). *Dicţionar de cuvinte derivate*, Editura Herra, Bucureşti.

Duško, V., Krstev, C. (2005). Derivational Morphology in a E-Dictionary of Serbian, In Zygmunt Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference*, Poznan, Poland, pp. 139-143.

Iordan, I. (1970). *Limba română contemporană*. Bucureşti, pp. 66-99.

Petic, M. (2010). Automatic derivational morphology contribution to Romanian lexical acquisition. *Special issue: Natural Language Processing and its Application. Research in Computing Science,* Mexico, vol. 46, pp. 67-78.

Santana, O., Perez J., Carreras, F., Rodriges, G. (2004). Suffixal and Prefixal Morpholexical Relationships of Spanish. *Lecture Notes in Artificial Intelligence*, Ed. Springer-Verlag, pp. 407-418.

DEZVOLTAREA UNUI PARSER DE ROLURI SEMANTICE PENTRU LIMBA ROMÂNĂ

DIANA TRANDABĂȚ[1], DAN CRISTEA[1,2]

[1]*Institutul de Informatică Teoretică, Academia Română, Filiala Iași*
[2] *Universitatea „Al. I. Cuza", Facultatea de Informatică, Iași*

{dtrandabat, dcristea}@info.uaic.ro

Rezumat

Parsarea semantică, prin identificarea și clasificarea entităților semantice în context, precum și a relațiilor dintre ele, are un mare potențial pentru aplicații cum ar fi rezumare de text, sisteme de întrebare-răspuns, sau traducere automată. Astfel, prin dezvoltarea unui sistem care să adnoteze automat roluri semantice pentru limba română, lucrarea de față reprezintă un pas intermediar important pentru înțelegerea automată a limbajului natural. Pentru dezvoltarea parserului de roluri semantice pentru limba română, a fost întâi necesară dezvoltarea unui corpus de antrenare, adnotat la roluri semantice. Această resursă adnotată de roluri semantice pentru limba română a fost creată prin utilizarea unei metode de transfer automat al rolurilor semantice, plecându-se de la o resursă dezvoltată pentru limba engleză (FrameNet). Ulterior, folosind o platformă pentru dezvoltarea de sisteme de etichetare supervizată a rolurilor semantice (PASRL), au fost antrenați mai mulți algoritmi de învățare folosindu-se corpusul dezvoltat, iar modelul obținut cu cel mai bun algoritm a fost salvat. Sunt discutate rezultatele aplicării acestui model pe texte neadnotate din limba română.

1. Introducere

Semantica cadrelor (*Frame Semantic*s) (Fillmore, 1982) este o teorie lingvistică care descrie structura conceptuală ce stă la baza înțelesului lingvistic. Cadrul semantic reprezintă o structură scriptică de inferențe, legată prin convenții lingvistice de înțelesul unităților lexicale. Fiecare cadru identifică un set de constituenți, roluri semantice (*frame elements*) care îl definesc și o serie de unități lexicale (cuvinte) ce participă la actualizarea sa. Conceptul de *unitate lexicală* (sau cuvânt predicațional) este central pentru resursele de cadre semantice. Unitatea lexicală este acel cuvânt sau sens al unui cuvânt polisemantic pentru care se definesc proprietățile combinatorice, definit printr-o lemă, o parte de vorbire și un cadru. Descrierea în termeni de semantică a cadrelor a unei unități lexicale identifică cadrele care formează un înțeles dat și specifică modul în care rolurile semantice sunt realizate în interiorul unor structuri dominate de cuvântul țintă.

Sistemul prezentat este dedicat prelucrării limbajului natural plecând de la text neadnotat, cu scopul de a determina informațiile de natură semantică pentru verbele și substantivele predicaționale. Cuvintele predicaționale au o anumită valență, și anume capacitate combinatorie, însușire de a deschide anumite poziții libere care sunt ocupate de termenii învecinați în propoziție. Fiind strâns legată de semantica verbului, valența determină numărul și caracteristica funcțional-semantică și gramaticală a elementelor cerute de verb (constituenți esențiali și facultativi). Cel mai activ și mai important sub

aspectul valenței este verbul, însă pot exista și substantive sau adjective ce au trăsătura de predicaționalitate care realizează nucleul semantic al propoziției. Sistemul de adnotare automată a rolurilor semantice pentru limba română se bazează pe o resursă de cadre semantice creată pentru limba română prin transferul rolurilor din limba engleză (Trandabăț, 2007).

Lucrarea este structurată în 4 secțiuni. După o scurtă introducere în domeniul rolurilor semantice, secțiunea 3 începe cu descrierea modului de creare a unui corpus de antrenare adnotat la roluri semantice pentru limba română. Folosirea acestui corpus pentru dezvoltarea unui sistem de adnotare automată a rolurilor semantice este prezentată în secțiunea 3.2., iar secțiunea 3.3. prezintă succint modulele de pre-procesare necesare pentru folosirea parserului de roluri semantice pentru limba română. În secțiunea 4 sunt discutate direcții viitoare de dezvoltare și utilizare a sistemului de etichetare a rolurilor semantice pentru limba română.

2. Rolurile semantice și adnotarea lor automată

În procesul comunicării, cuvintelor predicaționale sunt completate cu constituenți adiționali. Gradul de necesitate al constituenților pentru plenitudinea semantică a enunțului alcătuit de predicat este însă diferit. De exemplu, verbul *a depinde* cere cu precădere un *actant* și un *obiect* (*cineva depinde de ceva*). Ceilalți determinați care pot apărea pe lângă verbul *a depinde* sunt actualizați explicit numai atunci când sunt ceruți de situația de comunicare. Astfel, unele dintre pozițiile sintactice descrise de verb pot rămâne libere, fără ca enunțul să aibă de suferit și, dimpotrivă, suprimarea sau neexprimarea altora poate avea drept efect generarea unor structuri incorecte, incomplete sub aspect structural și semantic. Din acest punct de vedere se disting două tipuri de valențe (și deci tipuri de roluri semantice): valențe esențiale, numite și *argumente* și valențe facultative, numite și *adjuncți*. Un rol semantic esențial (numit și rol nucleu) este un rol care instanțiază o componentă necesară din punct de vedere conceptual pentru definirea și diferențierea cadrului. *Cumpărător*, *Vânzător*, *Bani* și *Bunuri* sunt exemple de roluri nucleu pentru cadrul semantic *Comerț*. Rolurile semantice care introduc evenimente independente sau distincte, adiționale evenimentului principal, sunt facultative, și se mai numesc și roluri periferice. Rolurile periferice marchează dimensiunea temporală, spațială, modală etc., dar nu caracterizează în mod unic cadrul și pot actualiza orice cadru semantic evenimențial. *Mod*, *Mijloace*, *Scop*, *Rată* și *Unitate* sunt exemple de roluri periferice pentru cadru semantic *Comerț*.

Etichetarea rolurilor semantice este o component importantă a înțelegerii limbii, și a fost considerată de mai multe sisteme computaționale. Sistemele tradiționale de parsare și înțelegere, inclusiv implementări bazate pe gramatici de unificare, se bazează pe gramatici dezvoltate manual, care trebuie să anticipeze fiecare mod în care rolurile semantice ar putea fi realizate sintactic. Scrierea acestor gramatici este consumatoare de timp, și de obicei, astfel de sisteme au o performanță limitată. Metodele bazate pe învățare promit o generalizare dincolo de numărul relativ mic de instanțe sau roluri considerate. Astfel, diverse strategii de învățare au fost folosite pentru adnotarea automată a rolurilor semantice: estimarea probabilităților (Gildea and Jurafsky, 2002),

arbori de decizie (Surdeanu et al., 2003), maşini cu suport vectorial (Pradhan et al., 2005) şi învăţarea bazată pe memorie (Morante et al., 2008).

Un dezavantaj important al celor trei sisteme prezentate este că acestea nu tratează predicatele nominale, fiind construite doar pentru predicate verbale[1]. Mai mult, acestea iau în considerare doar un singur predicat pentru fiecare propoziţie, chiar dacă nu acesta este întotdeauna cazul. De exemplu, în propoziţia:

Acordarea premiului Nobel Preşedintelui Obama a fost dezbătută pe larg.

avem două cuvinte predicaţionale, `acordarea`, având ca *Temă* `premiului Nobel`, şi `dezbătute`, având două roluri, o *Temă* `acordarea Premiului Nobel Preşedintelui Obama` şi un adjunct *Modal* reprezentat de grupul prepoziţional `pe larg`.

Sistemul prezentat în lucrarea de faţă tratează atât verbele, cât şi substantivele predicaţionale pentru limba română.

3. Dezvoltarea parserului de roluri semantice pentru limba română

3.1 Corpusul de antrenare

Pentru dezvoltarea resurselor de cadre semantice pentru limbile spaniolă, germană şi japoneză, s-a plecat de la un corpus specific fiecăreia dintre aceste limbi, adnotat manual la roluri semantice. În această secţiune descriem o metodă de constituire a unui corpus românesc de cadre semantice prin import din limba engleză. Premisa programului de importare automată a rolurilor semantice din limba engleză pentru limba română se bazează pe proprietatea cadrelor semantice de a exprima concepte la nivelul structurii de adâncime, valabile pentru toate limbile, actualizarea sintactică având loc ulterior la nivelul structurii de suprafaţă, diferit pentru fiecare limbă în funcţie de constrângerile sintactice şi morfologice. Programul de transfer automat (Trandabăţ, 2007) are la bază corelarea rolurilor semantice exprimate în limba engleză cu traducerea pentru limba română a cuvintelor ce realizează rolurile respective. Ulterior, aceste adnotări automate au fost validate manual. Această abordare este similară celei din (Barbu Mititelu and Ion, 2005) folosită pentru transferul relaţiilor de dependenţă verbală dintr-un corpus aliniat englez-român.

Astfel, folosind fişierele XML ale propoziţiilor engleze adnotate la roluri semantice, se creează automat un set de fişiere XML ce conţin un corpus de propoziţii adnotate la nivelul rolurilor semantice pentru limba română, din care urmează să se extragă cadrele semantice. Programul de adnotare foloseşte ca fişiere de intrare: **(i)** fişierele XML pentru unităţile lexicale engleze, care conţin propoziţii adnotate, şi **(ii)** fişierele cu alinierea propoziţiilor engleze - româneşti.

Pentru realizarea corpusului românesc au fost alese aleatoriu 110 de propoziţii din resursa FrameNet pentru limba engleză. Pe lângă aceste propoziţii, a fost ales cadrul semantic *Event* cu toate cadrele semantice corelate lui, totalizând alte 984 de propoziţii

[1] Începând cu competiţia ConLL2008, predicatele nominale au fost introduse în sistemele de identificare a rolurilor semantice, împreună cu predicatele verbale.

adnotate. În plus, încă 400 de propoziţii, obţinute din competiţia de aliniere la nivel de cuvânt pentru limbile engleză-română din cadrul ACL 2003 şi 2005[2], au fost adnotate cu un parser de roluri semantice pentru limba engleză, şi ulterior validate.

După selectarea propoziţiilor s-a realizat traducerea manuală a enunţurilor (exceptând cele 400 de propoziţii deja traduse) încercându-se menţinerea pe cât posibil a părţii de vorbire din limba engleză, măcar pentru unitatea lexicală care determină cadrul semantic, pentru a mări precizia alinierii.

Alinierea propoziţiilor englezeşti cu cele româneşti a fost realizată folosindu-se aliniatorul dezvoltat de Institutul de Cercetări în Inteligenţă Artificială al Academiei Române (Tufiş et al., 2005).

Au fost identificate 9 tipuri de import al adnotării, descrise detaliat în (Trandabăţ, 2010). Validarea rezultatelor s-a bazat pe detectarea cazurilor când importul a eşuat, încercând să descopere dacă problemele s-au datorat traducerii sau particularităţilor semantice şi sintactice ale limbii române. Au fost găsite doar puţine erori de traducere, şi chiar şi în acele cazuri, înţelesul fusese păstrat şi rolurile semantice fiind corect atribuite.

Majoritatea propoziţiilor constituie un corpus corect din punct de vedere al adnotării. Cazurile speciale de neconcordanţe ţin fie de inconsecvenţe de adnotare, fie de diferenţe culturale sau lingvistice dintre cele două limbi. Problemele de import au fost prezentate în (Trandabăţ & Husarciuc, 2008) şi pot fi grupate în următoarele clase:

- Cazuri în care există mai multe adnotări posibile în engleză – acelaşi constituent poate avea roluri tematice diferite. În limba română pot exista ambele roluri sau doar unul. Programul de import recunoaşte un singur rol pentru fiecare constituent, şi anume primul care a fost adnotat. Aceste cazuri sunt marcate pentru validare manuală. De exemplu, importul automat pentru propoziţia din limba engleză:

Traditional methods require that [the animal]$_{Protagonist}$ [bleed]$_{Cause}$ to [death]$_{TARGET}$ [after having its throat cut]$_{Time/Cause}$

este:

Metodele tradiţionale cer ca [animalul]$_{Protagonist}$ [să sângereze]$_{Cause}$ până la [moarte]$_{TARGET}$ [după ce i s-a tăiat gâtul]$_{Time}$.

- Cazuri în care limba română are roluri semantice ce nu apar în limba engleză (de exemplu argumente externe) ca în:

[Quit]$_{TARGET}$ [smoking]$_{Process}$.

[Lăsaţi]$_{TARGET}$ –[vă]$_{Protagonist}$ [de fumat].

unde rolul *Protagonist* nu există în propoziţia engleză, dar el apare în traducerea în limba română şi ar trebui adnotat, deşi programul de import nu poate transfera nimic.

- Cazuri în care un rol din limba engleză nu este exprimat în limba română deoarece nu este exprimat explicit în structura sintactică (de obicei este vorba de subiect care este

[2] Propoziţiile au fost descărcate de pe pagina Radei Mihalcea: http://www.cse.unt.edu/~rada/downloads.html #romanian

obligatoriu în limba engleză, dar poate lipsi fără a invalida enunțul în limba română). Un exemplu în care rolul *Manner* este inclus în cuvântul țintă este:

[Blood]$_{Undergoer}$ [had congealed]$_{TARGET}$ [thickly]$_{Manner}$ [on the end of the smashed fibula]$_{Place}$.

[Sângele]$_{Undergoer}$ [se îngroșă]$_{TARGET}$ [spre capătul fibulei zdrobite]$_{Place}$.

- Diferențe privind modul de formulare a enunțurilor în cele două limbi. Cele mai multe exemple de acest fel găsite au fost datorate folosirii în limba engleză a verbelor copulative, modale, sau a verbelor support, traduse în limba română fără folosirea unui astfel de verb. Astfel, cadrele semantice din cele două limbi sunt diferite, iar rolurile din limba engleză, deși importate corect, pot fi inexistente în cadrul semantic al verbului din limba română.

And I was surprised at how easily [my eyes]$_{Entity}$ [became]$_{TARGET}$ [accustomed to seeing]$_{Final\_state}$ in the light of the head torch .

Și am rămas surprins cât de ușor [s- au obișnuit]$_{TARGET}$ [ochii mei]$_{Entity}$ [cu vederea]$_{Final\_state}$ la lumina făcliei principale .

Adnotările automate ale rolurilor semantice pe corpusul românesc au fost verificate pentru a se extrage situațiile de neconcordanță. O dezvoltare ulterioară este realizarea, plecând de la cazurile de diferențe lingvistice, de reguli automate ce vor fi implementate pentru îmbunătățirea rezultatelor programului de transfer automat al rolurilor semantice.

În plus față de adnotarea semantică, corpusul a fost îmbogățit cu adnotare la partea de vorbire, folosindu-se serviciul web RACAI (Tufiș et al., 2008), și adnotare a dependențelor sintactice, folosindu-se un parser creat prin antrenarea MALTParser (Nivre, 2003) pe un corpus de fraze românești adnotate manual cu dependențe FDG.

3.2 PASRL

Folosind resursa descrisă în secțiunea precedentă, a fost dezvoltat un sistem de creare a unui modul de adnotare automată a rolurilor semantice, numit PASRL (Platform for Adjustable Semantic Role Labeling), descris în (Trandabăț, 2010). Similar cu arhitectura generală a sistemelor de etichetare a rolurilor semantică (Marquez et al., 2008), PASRL este compus din două sub-sisteme principale: un modul de predicție a predicatului și un modul de predicție a argumentelor. Modulul de predicție a predicatului are la rândul lui două configurații posibile, corespunzând identificării predicatelor și a sensurilor acestor predicate succesiv sau simultan (figura 1).

Identificarea predicatelor – acest modul primește ca intrare propoziția analizată sintactic (cu informații despre părțile de vorbire și dependențele sintactice) și decide care dintre verbele și substantivele din propoziție sunt predicaționale, și deci pentru care trebuie căutate roluri semantice;

Identificarea sensului predicatelor – după identificarea predicatelor dintr-o propoziție, trebuie stabilit sensul fiecărui predicat (dintre sensurile din corpusul de antrenare), deoarece sensuri diferite pot cere o structură diferită de roluri semantice;

Identificarea rolurilor semantice – acest modul identifică, pentru fiecare predicat, ce rol are fiecare dependent al predicatului.

Pentru fiecare sub-problemă (nivel din figura 1), modulele au trei variante, legate de dimensiunea corpusului de antrenare: folosind toate datele de test, folosind numai datele care au drept cuvânt predicațional un substantiv, respectiv verb (NP / VP), sau folosind numai datele corespunzând unui anumit tip de rol. Pentru fiecare modul din figura 1, sunt antrenați un set de algoritmi de clasificare din Weka (Witten and Frank, 2005). După rularea tuturor clasificatorilor pentru toate modulele, performanța lor este comparată, iar calea care obține cea mai mare performanță este considerată cea mai bună configurație. Un exemplu de astfel de configurație poate fi: Rularea modelului creat pentru identificarea predicatelor folosind tot corpusul de antrenare și drept algoritm de clasificare arborii de decizie, urmat de modelul creat pentru identificarea sensului predicatului separat pe verbe/substantive (NP / VP) create tot folosind arbori de decizie, și de modelul de identificare a argumentelor pentru fiecare tip de argument în parte, creat folosind algoritmul Naive Bayes. Modelele pentru cea mai bună configurație sunt salvate iar configurația este folosită ulterior pentru a adnota texte noi cu roluri semantice.

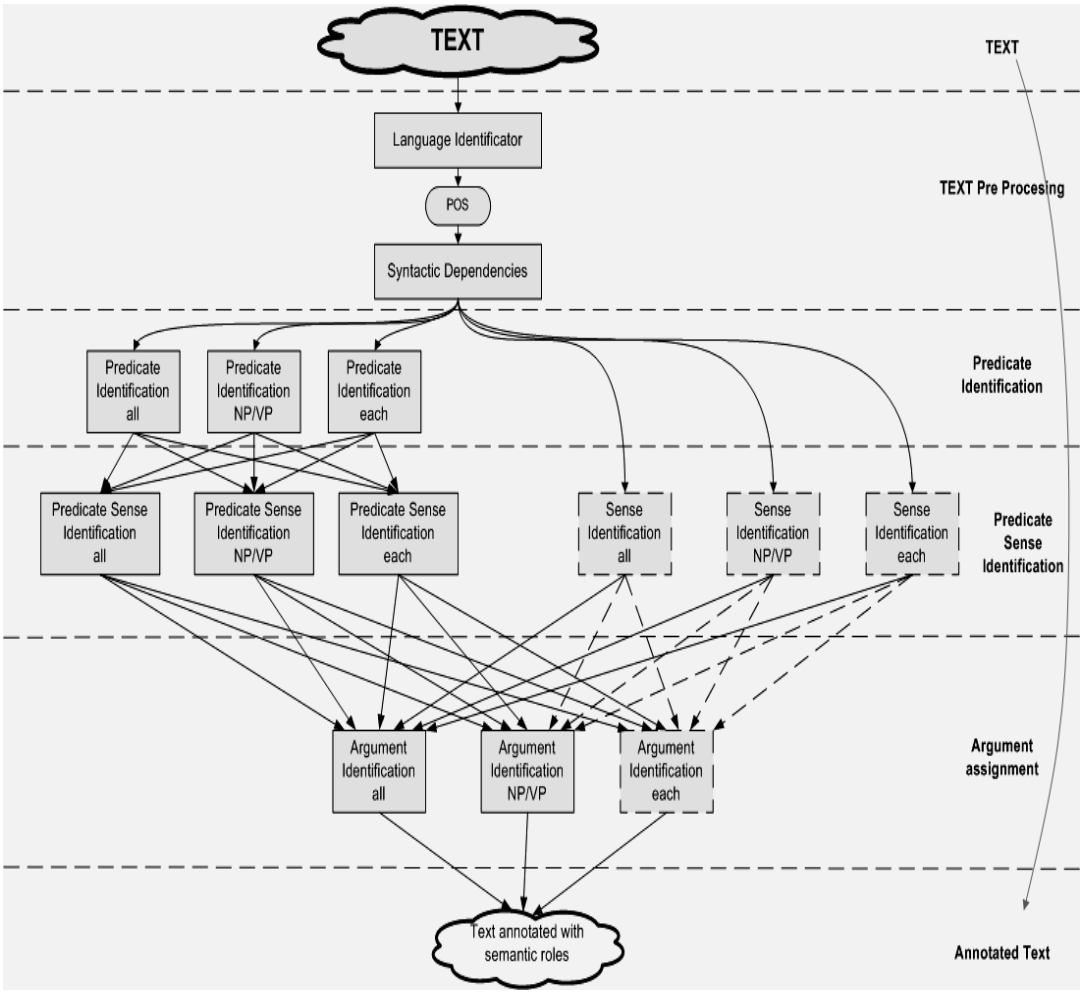


Figura 1: Arhitectura sistemului PASRL

Pentru stabilirea instanțelor, pentru fiecare predicat sunt considerați dependenții lui sintactici și contextul său imediat (5 cuvinte înainte și după predicat). Trăsăturile folosite în modulului de identificare a predicatului și pentru modulul de identificare a sensului predicatului sunt extrase din resursa creată pentru limba română:

1. o trăsătură binară care stabilește dacă cuvântul predicațional se află sau nu în lista de predicate adnotate în resursa pentru limba română[3];
2. numărul de dependenți sintactici (dacă un cuvânt nu are dependenți, atunci nu contează dacă el este sau nu predicațional, pentru că nu este considerat predicat pentru nici un constituent din propoziție[4]);
3. numărul de dependenți sintactici situați înainte/după predicat (de ex., mai mulți dependenți înainte de predicat indică o topică anormală sau un verb la diateza pasivă, deci rolurile semantice pot fi inversate);
4. părțile de vorbire și grupurile din care fac parte *n* cuvinte înainte, respectiv după cuvântul predicat, în ordinea de suprafață a propoziției, unde *n* are valoarea 5, stabilită empiric, dar poate fi modificată dacă este dată ca parametru;
5. relațiile de dependențe sintactice a celor *n* cuvinte;
6. trăsătură binară care stabilește dacă părintele dependentului sintactic este predicatul sau nu;
7. trăsătură binară care stabilește dacă părintele dependentului sintactic este unul dintre cele *n* cuvinte;
8. trăsătură binară care stabilește dacă predicatul are dependenți care nu se găsesc între cele *n* cuvinte.

Pentru modulul de identificare și clasificare a rolurilor semantice, instanțele sunt calculate considerându-se toate cuvintele din propoziție și următoarele trăsături:

1. partea de vorbire a cuvântului;
2. dependența sintactică față de predicat;
3. grupul sintactic din care face parte;
4. partea de vorbire a regentului cuvântului;
5. distanța de la cuvânt până la predicat, înainte sau după (în număr de cuvinte);
6. trăsătură binară stabilind dacă cuvântul investigat este punctuație sau nu;
7. trăsătură binară stabilind dacă cuvântul investigat este fiu al predicatului sau nu;
8. trăsătură binară stabilind dacă regentul cuvântului investigat este „`a fi`”, un verb copulativ sau modal;
9. hipernimul din WordNetul românesc (Tufiș et al., 2006), pentru substantive. Hipernimele sunt clasificate în prima clasă găsită, mergând bottom-up în ierarhia hipernimelor, dintre: spațial, temporal, cantitate, obiect, persoană, entitate.

PASRL a fost rulat pentru limba română, și cea mai bună configurație a fost: identificarea predicatelor cu arbori de decizie folosind tot corpusul, identificarea sensurilor predicatelor folosind algoritmul SimpleCart antrenat separat pentru predicate verbale, respective nominale, urmat de identificarea argumentelor folosind arbori de

[3] Resursa de roluri pentru limba română permite crearea unei liste a predicatelor (și a sensurilor acestora) pentru care există adnotări. Această listă este momentan restrânsă numeric, dar o dată cu îmbogățirea resursei românești, ea va fi actualizată, această trăsătură devenind mai importantă pentru determinarea acelor cuvinte care pot avea roluri.

[4] În această versiune a parserului, nu sunt adnotate rolurile implicite sau cele neexprimate.

decizie. Evaluarea performanţelor folosind 10-fold cross-validation pe corpusul de antrenament indică o precizie de 74% şi un recall de 64%.

3.3 Ajustarea parserului pentru text neadnotat

Pentru a adnota un text nou în limba română cu roluri semantice, utilizând modele salvate de PASRL, informaţiile sintactice sunt adăugate folosind serviciile web RACAI, iar relaţiile de dependenţă folosind un model antrenat pentru limba română cu parserul MALT. Evaluările preliminare indică o F-measure de 54%, ceea ce indică un început promiţător, având în vedere dimensiunea redusă a corpusului pe care s-a efectuat învăţarea (1500 de propoziţii).

4. Concluzii şi direcţii viitoare

Intuiţia conform căreia analiza semantică poate reprezenta o contribuţie pozitivă majoră pentru aplicaţiile de prelucrarea limbajului natural a motivat dezvoltarea unui număr de resurse de semantică lexicală care cuprind descrierea cadrelor semantice ale fiecărui verb dintr-o limbă. Cele mai importante astfel de proiecte sunt PropBank (Palmer et al., 2005) şi FrameNet (Baker et al., 1998). Valoarea acestor resurse este direct dependentă de cantitatea de informaţii pe care le conţin, lucru de asemenea dependent de nivelul efortului implicat în dezvoltarea lor. Această lucrare prezintă un corpus românesc de roluri semantice realizat printr-o metodă de transfer şi un sistem de adnotare a rolurilor semantice pentru limba română. Una din principalele direcţii urmărite în viitor este îmbunătăţirea resursei de roluri semantice (şi implicit şi a sistemului de etichetare automată) prin transferarea altor propoziţii din limba engleză. De asemenea, metoda de import poate fi folosită şi pentru alte domenii semantice. În această direcţie, vom investiga transferarea adnotărilor referitoare la referinţele anaforice, urmând testele din (Postolache et al., 2006).

Sistemul de etichetare a rolurilor semantice pentru limba română va fi în curând disponibil sub formă de serviciu web, pentru a putea fi implicat în dezvoltarea sistemelor aplicative de prelucrare a limbajului natural pentru limba română, cum sunt sistemele de rezumare automată (Cristea et al., 2005) sau de întrebare-răspuns (Iftene et al., 2009), dar şi pentru a fi folosite în proiecte de integrare a limbii române în reţele europene (Cristea şi Pistol, 2009).

Mulţumiri. Autorii mulţumesc colectivului Institutului de Cercetări în Inteligenţa Artificială al Academiei Române pentru accesul public la serviciile web. De asemenea, mulţumim colegilor Augusto Perez şi Alex Moruz pentru adnotarea propoziţiilor FDG, respectiv antrenarea parserului de dependenţe sintactice pentru limba română. Cercetarea prezentată în această lucrare a fost finanţată de către Programul Operaţional Sectorial Dezvoltarea Resurselor Umane prin proiectul „Dezvoltarea capacităţii de inovare şi creşterea impactului cercetării prin programe post-doctorale POSDRU/89/1.5/S/49944.

Referințe bibliografice

Baker, C. F., Fillmore, C. J., and John B., L. (1998). *The Berkeley FrameNet project*. In Proceedings of the COLING-ACL, Montreal, Canada.

Barbu Mititelu, V. and Ion, R. (2005). *Automatic Import of Verbal Syntactic Relations Using Parallel Corpora*. In Proc. of the International Conference Recent Advances in Natural Language Processing, pages 329-333, Borovets, Bulgaria.

Cristea, D., Postolache, O., and Pistol, I. (2005). *Summarisation through discourse structure*. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005, volume 3406 of LNCS, pages 632-644. Mexico City, Mexico.

Cristea, D., and Pistol, I. (2009). *The Romanian language from Clarin perspective* (in Romanian). In Diana Trandabăț, Dan Cristea, and Dan Tufiș, editors, Proceedings of the Workshop „Linguistic Resources and Instruments for Romanian Language Processing", pages 55-64, Iasi, Romania.

Fillmore, C. J. (1982). *Frame semantics*, în Linguistics in the Morning Calm, Hanshin Publishing, Seoul , 111-137.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245-288.

Iftene, A., Trandabăț, D., Pistol, I., Moruz, M. A., Husarciuc, M., Cristea, D. (2009). *UAIC Participation at QA@CLEF2008*. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Penas, and Vivien Petras, (eds), Evaluating Systems for Multilingual and Multimodal Information Access, 9th CLEF 2008, Revised Selected Papers, vol 5706 of *LNCS*, pg 385-392. Springer.

Marquez, L., Carreras, X., Litkowski, K., C. and Stevenson, S. (2008). Semantic role labeling: An introduction to the Special Issue. *Computational Linguistics*, 34(2):145-159.

Morante, R., Daelemans, W., and Van Asch, V. (2008). *A combined memory-based semantic role labeler of English*. In Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp 208-212, Manchester, UK.

Nivre, J. (2003) *An efficient algorithm for projective dependency parsing*. In Proc. of the 8th International Workshop on Parsing Technologies (IWPT 03), pp. 149-160.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71-106.

Postolache, O., Cristea, D. and Orasan, C. (2006). *Transferring Coreference Chains through WordAlignment*. In Proceedings of the 5th Language Resources and Evaluation Conference (LREC2006), Genoa, Italy, May.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H., and Jurafsky, D. (2005). Support vector learning for semantic argument classification. *Machine Learning Journal*, 60(13):11-39.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003) *Using predicate-argument structures for information extraction*. In Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics, pages 8-15, Tokyo.

Trandabăţ, D. and Husarciuc, M. (2008). *Romanian semantic role resource*. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may.

Trandabăţ, D. (2010). *Natural language processing using semantic frames*. PhD Thesis, http://students.info.uaic.ro/~dtrandabat/thesis.pdf.

Trandabăţ, D. (2007). *Semantic frames in Romanian natural language processing systems*. In Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 29-32, Rochester, New York. Association for Computational Linguistics.

Tufiş, D., Ion R., Ceauşu, Al., Ştefănescu, D. (2005) *Combined Aligners* in Proceeding of the ACL2005 Workshop on „Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, June, 2005

Tufiş, D., Barbu Mititelu, V., Bozianu, L., and Mihaila, C. (2006). *Romanian wordnet: New developments and applications*. In Proceedings of the 3rd Conference of the Global WordNet Association, pages 337-344, Seogwipo, Jeju, Republic of Korea, Jan.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008). RACAI's Linguistic Web Services, în Proceedings of LREC 2008 (Language Resources and Evaluation Conference), May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.

SUPRAVEGHEREA PE INTERNET: PĂREREA CONSUMATORILOR DESPRE DIVERSE PRODUSE SAU EVENIMENTE

ADRIAN IFTENE, ALINA-ELENA MIHĂILĂ, GEORGE ALEXANDRU VLAD, GETA STANCU

Universitatea „Al.I.Cuza", Facultatea de Informatică, Iaşi – România

{adiftene, elena.mihaila, george.vlad, geta.stancu}@info.uaic.ro

Rezumat

Lucrarea de faţă abordează şi încearcă să ofere o soluţie viabilă la una din problemele zilelor noastre: supravegherea şi extragerea de informaţii relevante de pe Internet. Volumul foarte mare al informaţiilor existente pe siteurile web, pe forumuri sau pe pagini personale, fac ca găsirea informaţiilor care ne interesează să fie complicată şi consumatoare de timp. În plus datorită succesului lor, reţele sociale precum Twitter, MySpace, Facebook, Flickr au un număr de membri cu interese comune în creştere, iar informaţiile existente aici au şi ele un volum din ce în ce mai mare. Într-o astfel de reţea socială, pornind de la o temă de discuţie, utilizatorii îşi exprimă liber părerile, adaugă link-uri sau poze relevante.

Sistemul pe care l-am construit se foloseşte de informaţiile existente pe Internet şi oferă persoanelor care-l folosesc o modalitate uşoară de a afla opinii pozitive sau negative asupra unui subiect de discuţie. Informaţiile sunt căutate pe paginile web (unde preferăm blog-urile personale, forumurile sau comentariile utilizatorilor) şi în reţeaua socială Twitter. Identificarea şi clasificarea opiniilor se face prin identificarea unor declanşatori de emoţii şi prin calcularea unor valenţe asociate contextelor din care acestea fac parte.

1. Introducere

În ultimii ani am remarcat o explozie a comunităţilor de utilizatori în cadrul reţelelor sociale, a forumurilor, sau a site-urilor dedicate unor categorii de persoane care folosesc Internetul pentru a face schimb de mesaje, documente, link-uri, etc. În cadrul acestor comunităţi, aceşti participanţi îşi exprimă liber opiniile în legătură cu subiecte de interes comune, critică sau laudă anumite aspecte ale temelor comune de discuţie. Interesant este faptul că, în cazul unui produs, de exemplu, pe baza acestor impresii exprimate liber de utilizatori, ne putem face o părere despre calitatea acestuia, putem identifica avantajele folosirii sale şi chiar punctele slabe ale acestuia.

De exemplu, site-ul *tweetfeel*[1] vă permite să căutaţi pe Twitter[2] ultimele mesaje ale utilizatorilor de aici, în legătură cu un subiect care vă interesează. Aceste postări sunt clasificate în două categorii: prima cu comentarii negative şi cea de a doua cu comentarii pozitive pe baza unor cuvinte cheie care sunt identificate în mesajele utilizatorilor. Clasificarea într-una din cele două categorii se face pe baza distanţei între cuvintele cheie care caracterizează cele două categorii şi cuvintele pe care le căutăm pe

[1] Tweetfeel: http://www.tweetfeel.com/#iphone
[2] Twitter: http://twitter.com/

Twitter, fără a se ţine cont de contextul în care apar acestea. Deoarece această măsură nu ţine cont de semantica textului pentru care se calculează, nu întotdeauna clasificarea comentariilor se face în categoria corectă. Cu toate acestea, contextele în care apar cuvintele cheie folosite de utilizatori în procesul de căutare sunt de ajutor în crearea unei opinii cu privire la ceea ce-l interesează pe acesta.

Lucrarea de faţă prezintă în prima parte câteva noţiuni legate de ceea ce reprezintă supravegherea utilizatorilor şi a informaţiilor, şi apoi câteva din componentele de bază ale sistemului creat de noi. Aplicaţia creată are scopul de a afla părerea consumatorilor în legătură cu anumite produse sau evenimente care au avut loc. În cea de a doua parte a lucrării avem prezentate două studii de caz pentru a arăta modul în care funcţionează aplicaţia noastră. În primul, arătăm cum putem afla opinia consumatorilor cu privire la anumite obiceiuri legate de sărbătoarea Paştelui şi în cel de al doilea, cum au privit românii, în comparaţie cu americanii acordarea premiului Nobel preşedintelui american, Barrack Obama.

2. Supravegherea utilizatorilor şi a informaţiilor de pe Internet

Foucault consideră că în prezent societatea noastră se comportă ca o societate de supraveghere şi ca o societate disciplinară. În această societate „*individul este atent fabricat în ea, în conformitate cu o întreagă tehnică de forţe şi corpuri*" (Foucault, 1977).

Giddens, precizează că la baza supravegherii stă acumularea de informaţii care pot fi stocate de o agenţie sau o colectivitate, precum şi supravegherea activităţilor subordonaţilor de către superiorii lor (Giddens, 1981). Statul naţional modern va fi de la început o societate informată, deoarece se ocupă de colectarea şi păstrarea informaţii referitoare la cetăţenii săi (nume, adresă, buletin, paşaport, zile de naştere, căsătorii, decese, statistici demografice şi fiscale, etc.) în scopul organizării administrării.

Conform (Fuchs, 2009) supravegherea pe Internet este legată de supravegherea informaţiile care circulă pe Internet şi se face din mai multe motive, din care cele mai importante din ziua de azi sunt legate de securitatea naţională (mai ales după atentatul de la 11 Septembrie) şi de interese comerciale ale marilor firme. De exemplu, o aplicaţie care poate extrage informaţii de pe Internet, legate de acţiunile unor grupuri de utilizatori, care doresc să protesteze cu ocazia unor întâlniri ale oamenilor politici, poate oferi o motivaţie puternică pentru constituirea unui grup mai mare care să asigure securitatea pe toată durata evenimentului.

Majoritatea tehnicilor ce presupun supravegherea pe Internet implică monitorizarea datelor şi a traficului pe Internet. Calculatoarele conectate la Internet comunică între ele cu ajutorul mesajelor, care sunt sparte în bucăţi mai mici, pentru a putea fi transmise. Aceste pachete sunt mai apoi transmise prin intermediul unei reţele de calculatoare, din nod în nod, până când îşi găsesc destinaţia. Aici, toate pachetele sunt asamblate la loc, formând mesajul iniţial. Urmărirea traiectoriei acestor pachete, precum şi a conţinutului acestora, ţine de supravegherea pe Internet.

Conceputul de supraveghere pe internet se află în opoziţie cu ideea de securitate a Internetului. Asta deoarece, potrivit Netlingo[3] avem că: „*Informaţiile care circulă pe Internet, de obicei, au un traseu ocolitor până la calculatorul de destinaţie, prin mai multe calculatoare intermediare. Traseu real nu este sub controlul celui care doreşte să realizeze un astfel de transfer. Prin urmare, la fiecare calculator intermediar există riscul ca cineva să urmărească atent datele care îl traversează şi să facă copii ale acestora. Un calculator intermediar poate induce în eroare pe oricine, acesta putând foarte uşor să-şi ascundă adevăratele sale intenţii. În acest mod acesta poate avea acces la informaţii confidenţiale, cum ar fi parole sau numere de card de credit*".

3. Prezentarea sistemului

Sistemul construit de noi foloseşte două modalităţi de extragere a informaţiilor dorite de pe Internet. Prima porneşte de la o interfaţă simplă în care utilizatorul poate introduce o întrebare în limbaj natural. Această întrebare este procesată cu ajutorul componentei de prelucrare a întrebării prezentate în (Iftene et al., 2009) şi din ea extragem cuvintele cheie după care face căutarea pe Internet. Această interfaţă folosind bibliotecile de căutare din Google API[4] (Google AJAX Search API[5]), extrage de pe Internet legăturile către siteurile cele mai relevante pentru căutarea noastră. În continuare, sistemul nostru apelează la o componentă Lucene Nutch[6] cu ajutorul căreia salvăm local conţinutul siteurilor şi apoi le indexează pe calculatorul pe care lucrăm. În final tot cu ajutorul componentei Nutch căutăm în index părerile utilizatorilor, pe care le clasificăm în comentarii pozitive şi în comentarii negative folosind un modul specializat în identificarea opiniilor şi sentimentelor în texte.

Cea de a doua este similară cu prima, numai că în loc să extragă informaţiile de pe Internet folosind Google AJAX Search API, foloseşte API-ul de la Twitter[7].

În continuare prezentăm componentele principale ale sistemului, fiecare cu principalele caracteristici.

3.1 Google AJAX Search API

Motoarele de căutare au început să facă parte din ce în ce mai mult din viaţa noastră de zi cu zi. Volumul foarte mare de informaţii care există pe Internet, face imposibilă căutare unor informaţii fără folosirea acestor componente. Folosirea unor astfel de module în cadrul aplicaţiilor pe care le construim, duce la mărirea eficienţei şi a dinamismului sistemului nostru, şi la o mai bună interacţiune cu informaţiile existente pe Internet.

Pentru a integra în aplicaţia noastră modulul de căutare Google am folosit biblioteci JavaScript existente în Google AJAX Search. Acestea ne permit să adăugăm funcţionalităţi noi aplicaţiei noastre, precum ar fi căutarea informaţiilor pe Internet.

[3] Netlingo: http://www.netlingo.com/
[4] Google API: http://code.google.com/
[5] Google AJAX Search API: http://code.google.com/apis/ajaxsearch/
[6] Lucene Nutch: http://lucene.apache.org/nutch/
[7] Twitter API: http://apiwiki.twitter.com/

În urma căutării realizate de utilizator, Google Search API ne întoarce o listă de legături către paginile cele mai relevante. Aceste pagini sunt mai apoi copiate local şi indexate cu ajutorul componentelor Nutch.

3.2 Apache Nutch

Nutch/Lucene este o platformă implementată în Java de tip open-source. Componenta Nutch a sistemului nostru este folosită atât pentru copierea locală a siteurilor web cât şi pentru indexarea acestora. După procesul de indexare, pe baza unei interfeţe web Nutch putem să efectuăm căutări în cadrul indexului creat.

În urma operaţiei de navigare şi preluare a informaţiilor de pe web sau din cadrul reţelelor locale, conţinutul siteurilor este salvat local într-o mulţime de documente denumită *Corpus*. În etapa imediat următoare, acest Corpus este indexat şi pregătit pentru a putea fi interogat de utilizatori. Ca în cazul celor mai multe motoare de căutare, componenta care se ocupă de partea de interogare este componenta cea mai complicată, dar şi cea mai importantă. Interogarea utilizatorului este transformată într-o mulţime de termeni index, care este trimisă mai apoi la motorul de interogare Nutch. Acesta va întoarce documentele cele mai relevante, adică cele în care se găsesc cele mai multe potriviri cu termenii din interogarea utilizatorului.

3.3 Twitter API

Twitter API este bazat în întregime pe protocolul HTTP. Metodele de preluare a datelor din Twitter API necesită o cerere de tip GET, iar metodele care fac introducere, modificare sau distrugere de date necesită cereri de tip POST. Sunt permise de asemenea cereri de tip DELETE, atunci când dorim să distrugem anumite informaţii.

Twitter API permite modificarea formatului extensiei cererii pentru a obţine rezultatele într-un alt format dorit de utilizator. Din formatele suportate în prezent de acest API (XML, JSON şi RSS sau Atom) aplicaţia construită de noi a folosit formatul XML.

Din funcţiile existente în acest API am folosit în primul rând funcţiile de căutare, care ne permit să extragem informaţiile dorite din postările utilizatorilor.

3.4 Identificarea opiniilor pozitive şi negative ale utilizatorilor

Identificarea înţelesului cuvintelor din texte a devenit în ultimii ani o direcţie importantă de cercetare în domeniul lingvisticii computaţionale. Abordările existente folosesc dezambiguizare, folosesc resurse lingvistice precum SensiWordNet sau încearcă identificarea declanşatorilor de emoţii (vezi lucrările (Hatzivassiloglou şi Mckeown, 1997), (Mihalcea et al., 2007) şi (Tufiş, 2009)).

Pentru identificarea opiniilor am folosit o metodă similară celei folosite în (Iftene şi Rotaru, 2010), care construieşte incremental o bază de date lexicală (care conţine cuvinte ce declanşează emoţii) similar cu (Balahur şi Montoyo, 2008). Elementele din această bază de date ne ajută să descoperim opiniile şi emoţiile în textele extrase de API-urile Google şi Twitter, iar în urma calculării unor valenţe globale ale textelor ne putem da seama dacă acestea reprezintă latura pozitivă, negativă sau neutră a unui sentiment.

Elementul de bază care identifică emoţiile este denumit *declanşator de emoţii* şi reprezintă un cuvânt sau un concept care poate oferi o interpretare emoţională a conţinutului textului. Iată câteva exemple de declanşatoare de emoţii: cuvinte precum „mândrie", „libertate", „stimă", „familie".

Baza lexicală de termeni a fost construită pornind de la 30 de termeni prezenţi în „piramida lui Maslow" (Maslow, 1943), care au fost traduşi în limba română. Adiţional am folosit WordNet-ul românesc (Tufiş et al., 2004) de unde am preluat sinonimele, antonimele, şi hiponimele.

După construirea bazei lexicale de termeni, următorul pas a fost de a atribui valenţe şi emoţii termenilor din această bază de date. Pentru aceasta am ţinut cont de următoarele reguli:

- Termenilor principali de declanşare a emoţiilor şi sinonimelor acestora li se acordă o valoare pozitivă.
- Termenilor hiponimi cu cei de mai sus şi termenilor derivaţi din termeni principali care sunt verbe li se acordă de asemenea o valoare pozitivă.
- Termenilor antonimi cu oricare din cei de mai sus li se acordă o valoare negativă.
- Valenţele oricăror termeni se modifică în funcţie de modificatorii (termeni care neagă, accentuează sau diminuează o valenţă) care îi însoţesc.

3.4.1 Modificatori de valenţă

Pentru a determina valenţele finale ale declanşatoarelor de emoţii se defineşte o mulţime de *modificatori de valenţă* (Balahur şi Montoyo, 2008). Un **modificator de valenţă** reprezintă un termen care modifică valenţa asociată unui alt termen pe care îl însoţeşte. Modificatorii pe care i-am identificat sunt de mai multe feluri:

- **Negaţii** – care modifică valenţa radical de la polul pozitiv la cel negativ sau invers. Aici modificatorii sunt reprezentaţi de cuvinte care introduc negaţia: „nu", „niciodată", etc.
- **Accentuatori** – care accentuează aspectul negativ sau pozitiv al unui declanşator. Din această mulţime fac parte adjective precum: „mare", „mult", „bine", „profund", „excepţional" sau adverbe care accentuează înţelesul întregului context din care fac parte „cu siguranţă", „sigur", „cert", „în definitiv".
- **Diminuatori** – care diminuează aspectul negativ sau pozitiv al unui declanşator ducându-l spre o valenţă neutră. Diminuatorii sunt reprezentaţi de adjective precum „mic", „puţin", „rău", „degrabă", de verbe modale: „a putea", „a fi posibil", „a trebui", „a vrea", de adverbe ca „posibil", „probabil". Verbele modale şi adverbele introduc noţiunea de incertitudine şi posibilitate şi diminuează valenţa emoţiei întregului context din care fac parte. De asemenea aceşti diminuatori ne ajută să facem distincţia între evenimente care au avut loc, ar fi putut să aibă loc, au loc în prezent sau vor avea loc în viitor.

De exemplu, cum ar suna un fragment în care există modificatori de valență diferiți: „*este minunat*" (forma inițială), „*nu este minunat*" (forma negată), „*ar putea fi minunat*" (forma diminuată), „*e absolut minunat*" (formă accentuată).

Există însă și termeni care, dacă sunt combinați/alăturați, produc sentimentul de ironie. Acesta este și cazul asocierii următoare: „*Excepționalul organizator a eșuat în rezolvarea problemei.*", unde termenul „excepțional" (are în mod uzual valență pozitivă) alături de termenul „a eșuat" (are valență negativă de obicei) ne duce cu gândul la faptul că „*organizatorul care era recunoscut pentru priceperea sa a dat-o în bară*".

În plus, pe parcursul testării și verificării aplicației, am identificat termeni noi specifici discuțiilor purtate pe forumuri, blog-uri sau rețele sociale și cu ajutorul lor am completat resursele obținute în pașii anteriori. Iată câteva exemple de termeni specifici: „brio", „super", „fine", „bună" sau succesiuni de caractere speciale care reprezintă iconițe emoționale: „:)" (față zâmbitoare ☺), „:(" (față tristă ☹), etc.

O observație importantă este următoarea: pentru ca un modificator de valență să își îndeplinească scopul (să modifice valența termenilor), trebuie ca în text să fi fost exprimată o atitudine (al cărei înțeles să poată fi modificat). De exemplu, în propoziția „Ion este acasă." care nu face decât să prezinte un fapt și nu o atitudine, introducerea unei negații „Ion nu este acasă" nu schimbă valența nici unui termen.

4. Studii de caz

În acest capitol am realizat două studii de caz, pentru a arăta modul de funcționare al sistemului nostru. Cu ajutorul acestui sistem utilizatorii vor avea toate aceste informații centralizate și în plus ele vor fi clasificate în comentarii pozitive și negative. Nu în ultimul rând aplicația, prin faptul că accesează simultan diferite zone de pe Internet și că oferă informațiile centralizat, îi va ajuta pe cei care o folosesc să reducă timpul necesar obținerii unor astfel de informații.

4.1 Sărbătoarea Paștelui

Primul studiu de caz este legat de sărbătoarea Paștelui. Scopul aplicației este de a veni în ajutorul celor care doresc să realizeze produse specifice acestei sărbători, ei neavând la dispoziție rețete disponibile și nici timp pentru a le căuta. Pentru acest studiu de caz am luat în considerare produsele tradiționale de Paște (cozonac, pască și ouă roșii) și am identificat opiniile utilizatorilor de pe Internet cu privire la anumite rețete sau modalități de preparare. Scopul final a fost de a identifica dacă putem găsi cu ajutorul aplicației o rețetă cât mai apropiată de gustul utilizatorului și care sunt avantajele și dezavantajele acesteia.

Pentru API-ul de la Google, pentru a fi siguri ca rezultatele întoarse vor fi relevante, am construit diverse interogări alegând diverse combinații ce folosesc cuvinte care reprezintă fie produse (*cozonac*, *pască*, *ouă roșii*), fie ingrediente ale acestora (*ciocolată*, *brânză*, *smântână*, *nucă*) (cu sau fără diacritice) și în plus am cerut ca aceste informații să se găsească pe pagini ce respectă formate de *forum* sau *blog*.

Pentru API-ul de Twitter am construit interogări mai simple (cu mai puţine cuvinte) pentru a avea şanse mai mari de reuşită folosind din nou cuvinte simple precum cele de mai sus. În final am realizat peste 20 de căutări pe Google şi 7 pe Twitter.

Pentru fiecare căutare realizată pe Google, Nutch-ul a preluat maxim 10 link-uri din rezultatele oferite pentru care a salvat local conţinutul site-urilor. Acest conţinut l-am procesat folosind componenta de identificare a emoţiilor şi am extras sentimentele pozitive şi negative.

În continuare vom vedea câteva exemple de propoziţii extrase cu ajutorul API-ului Google de pe forumuri în care am îngroşat cuvintele pentru care am identificat valenţe pozitive sau negative. Astfel, în urma căutărilor realizate pe tematici cu reţete de cozonaci, în comentariile utilizatorilor am găsit cuvinte declanşatoare de emoţii cu tentă pozitivă „se **pare** ca cozonacul mamei a trecut cu **brio** proba **:)**”, „Aluatul de cozonac e **foarte delicat** şi **sensibil**.”, „E o reţetă **super** simplă şi **super** gustoasă....”, şi cu tentă negativă „oricât aş vrea **nu** reuşesc sa fac acest cozonac”, etc.

Cu toate că în Twitter căutările au fost mult mai simple, rezultatele au fost considerabil mai puţine decât cele întoarse de motorul Google, deoarece am observat că numărul comentariilor în limba română sunt foarte puţine şi de cele mai multe ori conţin cuvinte şi chiar fragmente în limba engleză.

Iată câteva mesaje obţinute de pe Twitter în urmă căutărilor realizate pentru *cozonac*, *pască* şi *ouă roşii*: „deza@zozo_ro: **ciudată** combinaţie...dar suna gustos...acum mănânc din primul meu cozonac...si **ciudat**, a ieşit **bun**”, „richieTM: Deci cea mai **buna** plăcinta cu brânza e Pasca!!”, „lau_anca @RaduCeuca **super**! **dar** ai uitat de clasicele oua roşii **:)**”, etc.

În urma acestui studiu am observat că opiniile utilizatorilor de forum-uri şi blog-uri sunt mai consistente şi mai relevante decât cele ale utilizatorilor de pe Twitter, care sunt mai scurte şi care se folosesc de multe ori de imagini încărcate de aceştia.

4.2 Acordarea premiului Nobel lui Obama

Pe data de 10 octombrie, ştirea potrivit căreia Barack Obama a câştigat Premiul Nobel pentru Pace a ţinut prima pagină a ziarelor, atât în Romania cât şi în străinătate. Studiul de caz de faţă îşi propune să analizeze opiniile oamenilor vorbitori de limba română (şi pentru aceasta am căutat pe site-urile, blog-urile şi forum-urile scrise în limba română folosind interogări care folosesc cuvinte româneşti) şi a celor vorbitori de limba engleză (cu ajutorul interogărilor scrise în limba engleză) cu privire la aceasta decizie.

Rezultatele obţinute au arătat clar o diferenţă între comentariile celor din America (unde comentariile au fost în mare parte pozitive) şi celor din România (unde comentariile au fost în mare parte negative). Trebuie să precizăm că în America părerile au fost exprimate de persoane care au avut un rol important în alegerea lui Obama ca preşedinte, la mai puţin de un an de când aceştia l-au votat datorită discursurilor pe care acesta le-a ţinut în timpul campaniei sale. În acelaşi timp în România, părerile au fost exprimate de persoane care nu au fost implicate în nici un fel în alegerile din America.

Aplicaţia care foloseşte API-ul Google a extras informaţii de pe 20 de situri (zece din România şi zece din străinătate), însumând în total 327 de comentarii. Pentru a

identifica valenţele fragmentelor extrase pentru vorbitorii de limba engleză am tradus folosind Google Translate[8] resursele obţinute pentru limba română şi le-am extins folosind WordNet-ul[9] englezesc.

Cele 20 de situri au fost obţinute în urma folosirii unor interogări ce foloseau cuvinte cheie corespunzătoare unor întrebări ca mai jos:

1. Ce părere aveţi despre înmânarea premiului Nobel lui Barack Obama? (En: *What do you think about giving the Nobel Prize to Barack Obama*?)
2. Care sunt părerile pro şi contra la acordarea premiului Nobel lui Obama? (En: *What are the pros and cons to award the Nobel Prize Obama*?)
3. A meritat Obama premiul Nobel? (En: *Did Obama deserved the Nobel Prize*?)

În ceea ce priveşte siturile din România, o mare parte din cei care şi-au exprimat părerea au scos în evidenţă faptul că datorită acestei alegeri neinspirate premiile vor avea de suferit de acum încolo. Odată cu luarea acestei decizii, mulţi cred că premiile şi-au pierdut din importanţă şi că de acum câştigarea lor nu va mai fi o realizare atât de mare, întrucât un precedent a fost deja creat.

Iată câteva exemple de comentarii adăugate de utilizatori, în care am detectat o mare încărcătură emoţională corespunzătoare dezaprobării şi indignării (toate având aceeaşi sursă: http://economie.hotnews.ro/stiri-media_publicitate-6262253-presa-americana-radiosul-obama-accepta-premiul-pentru-pace-sau-cum-castigi-nobelul-12-zile.htm):

- „…e **culmea tupeului** sa dai premiul Nobel cuiva care era presedinte de doar 12 zile la data inchiderii perioadei de candidatura !!...”,
- „acest Nobel al lui Obama va ramane in istorie ca fiind **probabil** cel mai **nemeritat** din cate s-au acordat!”,
- „**Cred** ca prin aceasta **trista** alegere, nemeritata, Nobelul pt Pace e definitiv **compromis**.” , „a devenit evident ca **nu** are vreo legatura cu meritele sau rezultate reale obtinute. „.
- „**Cred** ca e o **jignire** la adresa celor care chiar fac ceva pentru pace in lume...”, „....pentru ce?!?! Pentru cele doua **razboaie** in care SUA sunt implicate?”
- „Un act de o **nesimtire** şi un **dispret incredibil** fata de sute de lideri politici”.

În siteurile în engleză (în principal pentru cele din America) am observat o altă atitudine. Deşi marea majoritate nu sunt de acord cu primirea acestei distincţii, criticile nu sunt îndreptate direct către preşedintele American, ci mai mult spre comisia care a acordat această distincţie.

În legătură cu declanşatoarele de emoţii, siturile în limba română conţin multe cuvinte cu valenţe negative şi modificatori care întăresc aceste valenţe (precum *incredibil*, *total*, *foarte*). În limba engleză se observă o tendinţă de a folosi cuvinte ce neutralizează valenţele (*poate că*, *posibil*, *probabil*).

[8] Google Translate: http://translate.google.com/
[9] WordNet: http://wordnetweb.princeton.edu/perl/webwn

În multe reacţii am observat o nedumerire generală („*nu înţeleg ce treaba are*", „*nu văd de ce*", „*nu ştiu de ce a fost ales*", „*nu cred că trebuia…*", „*nu consider că aceasta era cea mai bună alegere*") cu privire la luarea acestei decizii.

Puţinii oameni care nu comentează negativ această ştire, cu toate că nu sunt neapărat de acord cu luarea acestei decizii, îşi exprimă o admiraţie faţă de persoana în cauză.

5. Concluzii

Această lucrare prezintă principalele componente ale sistemului pe care-l oferim utilizatorilor care doresc să afle opinii pozitive sau negative asupra unor produse sau evenimente. Sistemul este o alternativă la căutarea clasică pe Internet, aducând nou posibilitatea de combinare a rezultatelor obţinute în urma căutărilor uzuale cu rezultate extrase de pe blog-urile utilizatorilor, forumuri sau chiar reţele sociale. Componentele principale ale sistemului se bazează pe componentele care ne ajută să căutăm şi să extragem informaţii relevante de pe Internet: API-urile de căutare de la Google şi de la Twitter.

Sistemul folosit de noi a folosit în calculul valenţelor asociate propoziţiilor distanţa între cuvintele cheie şi declanşatorii de emoţii, ţinând cont de modificatorii de valenţă. Din evaluările făcute, pe 6 teme de căutări efectuate, cu ajutorul a 3 persoane care au evaluat peste 100 de paragrafe extrase putem spune că suntem la început de drum, calitatea rezultatelor fiind încă modestă (în jur de 44%). Principalele probleme sunt datorate în principal faptului că declanşatorii de emoţii vizaţi nu se refereau la cuvintele cheie pentru care realizam căutarea. O altă problemă importantă este legata de faptul că prin modul de calcul al valenţelor asociate, preferăm propoziţiile scurte, care nu sunt întotdeauna relevante pentru ceea ce dorim să obţinem.

În continuare, dorim să implementăm în modul de calcul al valenţelor asociate elemente care să elimine principalele probleme prezentate mai sus, bazate pe identificarea rolurilor semantice. De asemenea dorim să realizăm o evaluare mai relevantă, cu circa 20 de teme de căutare şi cu mai mult de 10 evaluatori umani.

Mulţumiri. Cercetarea prezentată în această lucrare a fost finanţată de către Programul Operaţional Sectorial Dezvoltarea Resurselor Umane prin proiectul „Dezvoltarea capacităţii de inovare şi creşterea impactului cercetării prin programe post-doctorale POSDRU/89/1.5/S/49944.

Referinţe bibliografice

Balahur, A., Montoyo, A. (2008). *Applying a culture dependent emotion triggers database for text valence and emotion classification*. In journal Procesamiento del Lenguaje Natural, ISSN 1135-5948, Nº. 40. Pp. 107-114.

Foucault, M. (1977). *Discipline and punish*. New York: Vintage.

Fuchs, C. (2009). *Social Networking Sites and the Surveillance Society*. A Critical Case Study of the Usage of studiVZ, Facebook, and MySpace by Students in Salzburg in the Context of Electronic Surveillance Salzburg/Vienna, Austria. 2009.

Forschungsgruppe „Unified Theory of Information" - Verein zur Förderung der Integration der Informationswissenschaften. ISBN 978-3-200-01428-2.

Giddens, A. (1981). *A contemporary critique of Historical Materialism*. Vol. 1: Power, property and the state. London: Macmillan.

Hatzivassiloglou, V., Mckeown, K. R. (1997). *Predicting the semantic orientation of adjectives*. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Morristown. NJ, USA. pp. 174-181.

Iftene, A., Rotaru, A. (2010). *User Profile Modeling in eLearning using Sentiment Extraction from Text*. In Research in Computing Science, „Special issue: Natural Language Processing and its Applications", Vol.46, Pp.267-278, Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico 2010. ISSN: 1870-4069. Poster at 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICling 2010). 21-27 March. Iasi, Romania.

Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A.M., Husarciuc, M., Sterpu, M. and Turliuc, C. (2009). *Question Answering on English and Romanian Languages*. In Proceedings of the CLEF 2009 Workshop. 30 September - 2 October. Corfu, Greece.

Maslow, A. H. (1943). *A Theory of Human Motivation*. Psychological Review 50 (4). 370-96.

Mihalcea, R., Banea, C. şi Wiebe, J. (2007). *Learning Multilingual Subjective Language via Cross-Lingual Projections*. In Proceedings of the Association for Computational Linguistics (ACL 2007), Prague, June.

Tufiş, D. (2009). *Playing with Word Meanings*. In Lotfi A. Zadeh, Dan Tufiş, Florin Gh. Filip and Ioan Dziţac (eds.) From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence, Editing House of Romanian Academy. ISBN 978-973-27-1678-6, pp. 211-223.

Tufiş, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004). *The Romanian Wordnet*. Romanian Journal of Information Science and Technology, Volume 7, Numbers 1-2, pp. 107-124.

FOLOSIREA VERBELOR PENTRU DETERMINAREA INFERENȚELOR TEXTUALE

MIHAI ALEX MORUZ[1,2]

[1]*Universitatea „Al.I.Cuza”, Facultatea de Informatică, Iași – România*
[2]*Academia Română, Filiala Iași, Institutul de Informatică Teoretică – România*

mmoruz@info.uaic.ro

Rezumat

Această lucrare descrie o metodă de utilizare a verbelor și a structurii de argumente a acestora în vederea rezolvării de inferențe textuale pentru texte în limbile engleză și română. Analiza efectuată asupra unui set de perechi de inferențe din setul de test folosit la RTE-5 dovedește că semantica predicațională, așa cum a fost definită de Charles Fillmore și Beth Levin, este utilă pentru rezolvarea inferențelor textuale, întrucât 38% dintre perechile analizate sunt rezolvate direct folosind acest tip de semantică, iar 11% dintre perechi sunt rezolvate recurgând la elemente analiză verbală.

1. Introducere

Noțiunea de inferențe textuale (Textual Entailment în limba engleză) a fost descrisă de (Dagan, Glickman, 2004) ca o metodă de a generaliza noțiunea de variabilitate a limbajului natural în diverse probleme din domeniul limbajului natural, cum ar fi sistemele întrebare – răspuns, sumarizarea automată, etc. Conform (Dagan, Glickman, 2004), inferențele textuale sunt definite ca relație dintre un text coerent, **T**, și o expresie în limbaj natural, care este văzută ca o ipoteză, **H**. Spunem că **T** inferă **H**, (**H** este o consecință a lui **T**), notat **T⇒H**, dacă înțelesul lui **H**, interpretat în contextul lui **T**, poate fi dedus din înțelesul lui **T**. Relația de inferență definită mai sus este direcțională, deoarece deși un text poate fi dedus din un altul, reciproca nu este întotdeauna valabilă.

Definiția de mai sus, deși completă și corectă, este prea abstractă pentru a putea fi folosită direct în sisteme de procesare a limbajului natural. Recognising Textual Entailment Challenge (RTE) este principala campanie de evaluări pentru sistemele de rezolvare a inferențelor textuale, și a fost inițiată în 2005 de către PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning – http://www.pascal-network.org/) (Dagan et al., 2005), iar ultimele două ediții, cele din 2008 și 2009, au fost organizate în cadrul Text Analysis Conference (TAC – http://www.nist.gov/tac/). Din acest motiv au apărut o serie de adaptări ale definiției inițiale pentru inferențele textuale, care să poate fi utilizate în abordări practice.

Analizând diverse seturi de perechi de inferență, am ajuns la intuiția că majoritatea cazurilor pot fi rezolvate prin examinarea a două tipuri de informație:

- Relația dintre verbele din ipoteză și cele din text. Verbul trebuie considerat împreună cu argumentele și adjuncții săi; așadar, compararea dintre două verbe implică o comparare între două structuri complexe care sunt, în esență, propoziții atomice (clauze). Dacă verbele din **H**, împreună cu argumentele și adjuncții lor, sunt

susținute în **T**, rezultatul este ENTAILMENT; acest lucru se face prin identificarea acelor elemente din **T** care suportă informația din **H**.

- Fiecare argument sau adjunct este o entitate, care are atașat un set de proprietăți. Este posibil ca, în ciuda potrivirii la nivel de verbe, să apară diferențe în ceea ce privește proprietățile argumentelor sau adjuncților similari. Pentru a putea rezolva aceste cazuri, considerăm fiecare adjunct sau argument o entitate care are un set de atribute atașat, iar compararea se face atât pe nucleul entității cât și pe setul de atribute.

Această lucrare are ca scop principal determinarea relevanței verbelor pentru problema inferențelor textuale, așa cum au fost ele definite în cadrul acestei secțiuni. Mulțimile de proprietăți atașate argumentelor sau adjuncților sunt discutate pe scurt în secțiunea 4.3, cu referire la dezvoltări ulterioare.

Ideea folosirii verbelor pentru a rezolva inferențele textuale nu este nouă, și a fost utilizată de (Burchardt, Frank, 2006) și (Hickl, Bensley, 2007), (Hickl, 2008), printre alții. (Burchardt, Frank, 2006) aproximează inferențele textuale cu suprapunerea la nivel structural și semantic între text și ipoteză, și combină parsarea LFG cu semantica cadrelor, pentru a proiecta o reprezentare lexical semantică cu roluri semantice, și apoi calculează suprapunerea dintre **H** și **T**. (Hickl, Bensley, 2007), (Hickl, 2008) extrag o mulțime de propoziții acceptate universal, numite contexte discursive (*discourse commitments*); odată ce acestea au fost extrase, problema rezolvării inferențelor textuale se reduce la determinarea acordului contextelor discursive din **H** cu cele din **T**.

Diferența esențială dintre abordările precedente și soluția descrisă în cadrul acestei lucrări este aceea că soluția noastră folosește noțiunea de alternanțe verbale și rezultatele lui Beth Levin asupra claselor verbale (cunoscute drept *clase Levin*, și descrise în (Levin, 1993)). Mai precis, este determinată clasa Levin pentru fiecare verb din **H**, iar apoi se încearcă găsirea acelor verbe din **T** care au aceeași clasă Levin și care susțin ipoteza; dacă aceasta eșuează, se încearcă determinarea relațiilor dintre verbe pe baza descrierilor semantice atașate fiecărei clase. Această procedură a fost testată prin analiza detaliată a unui set de 200 de perechi de inferențe din datele de test din RTE-5[1], și am determinat că peste 38% dintre ele pot fi rezolvate corect folosind clase Levin. Pentru identificarea claselor Levin am folosit VerbNet[2], descris în detaliu în secțiunea 2.

O altă contribuție nouă a acestei abordări este gruparea proprietăților entităților în seturi de proprietăți; pentru două entități similare în **T** și **H**, inferența poate fi definită ca rezultatul comparării acestor seturi de atribute: ENTAILMENT înseamnă ca toate proprietățile din **H** sunt subsumate de proprietăți din **T**. Din cele 200 de perechi analizate, peste 29% pot fi rezolvate pe baza acestei metode.

2. Clasele Levin și VerbNet

Cea mai mare și mai frecvent utilizată clasificare a verbelor din limba engleză este cea descrisă în (Levin, 1993), mai larg cunoscută drept clase Levin. VerbNet (VN) (Kipper

[1] TAC 2009 Recognizing Textual Entailment Track, http://www.nist.gov/tac/2009/RTE/

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

et al., 2000; Kipper-Schuler, 2005) este un lexicon de verbe din limba engleză disponibil online care oferă descrieri sintactice şi semantice detaliate pentru clasele Levin, care sunt organizate într-o taxonomie mai rafinată. Conform (Kipper-Schuler et al., 2006) VN este un lexicon de verbe ierarhic, independent de domeniu, cu acoperire largă, care este aliniat cu resurse ca FrameNet (Baker et al., 1998) şi WordNet (Fellbaum 1998).

O clasă VerbNet este complet descrisă de o mulţime de verbe membre, rolurile tematice pentru structurile predicat-argument ale acestor verbe, restricţiile de selecţie pentru argumente şi cadre care conţin o descriere sintactică şi predicate semantice care au şi o funcţie temporală, similar cu descompunerea de evenimente propusă de (Moens, Steedman, 1988). VerbNet extinde clasele Levin iniţiale prin rafinarea acestora în subclase cu un mai mare grad de coerenţă sintactică şi semantică. Această ierarhie a fost extinsă, în urma studiului efectuat de (Korhonen, Briscoe, 2004), cu 57 de noi clase (de la aproximativ 200) şi 106 de noi alternanţe de diateză (de la aproximativ 200).

Întrucât această resursă este aliniată cu WN, ea poate fi uşor transferată pentru limba română, prin intermediul synseturilor, ceea ce ar permite aplicarea metodei descrisă în secţiunile 3 şi 4 pentru rezolvarea inferenţelor textuale în limba română.

2.1 Cadrele sintactice în VN

Fiecare clasă VN conţine un set de descrieri sintactice (cadre) care reprezintă posibile realizări de suprafaţă pentru structurile de argumente (tranzitive, intranzitive, cu grupuri prepoziţionale etc.) şi un număr de alternanţe de diateză, date de Levin ca parte a fiecărei clase (Kipper-Schuler, 2006). Fiecare cadru sintactic este descris folosind roluri tematice (*Agent*, *Patient* etc.), verbul şi toate celelalte elemente lexicale necesare pentru a construcţie sau alternanţă. Rolurile tematice pot fi restricţionate semantic (*animate*, *organization*, *human* etc.), iar cadrele sintactice pot fi şi ele restricţionate cu privire la prepoziţiile permise. Restricţii adiţionale pot fi impuse asupra rolurilor tematice, pentru a transmite natura sintactică a constituentului care este cel mai probabil asociat cu rolul tematic (*sentential*, *nominal* etc).

2.2 Predicate semantice

În afara descrierilor sintactice ataşate, fiecare clasă Levin are şi o descriere semantică, în forma unei conjuncţii de predicate semantice booleene precum *cause*, *motion*, sau *contact*. Fiecare dintre aceste predicate este asociat cu o variabilă eveniment *E*, care este folosită pentru specificarea momentului la care predicatul este adevărat. Relaţiile dintre verbe sau dintre clase de verbe (antonimie, consecinţă, etc) pot fi deduse pe baza acestor predicate semantice. În VN, aspectul unui verb este descris prin intermediul variabilei eveniment din predicate (Kipper-Schuler et al., 2006).

2.3 Statistici legate de VerbNet

În acest moment, VN versiunea 3.1, cea mai recentă versiune de VN, conţine descrieri pentru peste 5800 de sensuri verbale, distribuite în 270 de clase de nivelul întâi şi 200 de subclase. Descrierile acestor verbe folosesc 33 de roluri tematice, 36 de restricţii de selecţie, 296 de cadre principale (care pot fi extinse prin specificarea tipurilor de roluri tematice sau tipuri de complemente) şi 145 de predicate semantice. Lexiconul se

bazează și pe o ierarhie de prepoziții cu 66 de elemente. Acoperirea PropBank (Palmer et al., 2005) de către VN a crescut în această versiune la peste 90%. În plus, în VN 3.1 taxonomia definită de (Levin, 1993) a fost extinsă în profunzime, prin rafinarea unor clase existente și în lărgime, prin adăugarea claselor propuse de (Korhonen, Briscoe, 2004).

3. Algoritmul preliminar pentru compararea verbelor

Scopul studiului de fezabilitate, care va fi descris în detaliu în secțiunea 4, este acela de a testa dacă rezolvarea inferențelor textuale folosind relații între verbe, definite pe baza claselor Levin, este posibilă; rezultatele au arătat că acest fapt nu este doar posibil, ci că o astfel de abordare rezolvă direct peste 38% dintre perechile luate în considerare, și ajută, indirect, la rezolvarea a peste 11% altor perechi (o descriere detaliată este dată în secțiunea 4). Din acest motiv propunem un algoritm preliminar pentru rezolvarea inferențelor textuale pe baza relațiilor dintre verbe.

Algoritmul primește la intrare arbori sintactici pentru **T** și **H**, și întoarce rezultatul inferenței pentru perechea dată. Descrierea preliminară a algoritmului pentru rezolvarea de inferențe textuale pe baza relației dintre verbe este dată mai jos.

1. Se extrag clasele Levin pentru toate verbele din **T** și **H** și se atașează descrierea semantică corespunzătoare.
2. Se determină dacă fiecare verb din **H** are aceeași clasă Levin cu măcar un verb din **T**

   2.a. Dacă argumentele și adjuncții se potrivesc peste **T** și **H**, iar verbele nu sunt antonime, ENTAILMENT

   2.b. Altfel, dacă argumentele și adjuncții se potrivesc dar verbele sunt antonime, sau dacă argumentele nu se potrivesc (sunt entități conflictuale sau aceeași entitate dar cu atribute conflictuale), CONTRADICTION

   2.c. Altfel, UNKNOWN
3. Dacă nu există similaritate de clase Levin, se extrag relațiile dintre verbe pa baza descrierii semantice.

   3.a. Dacă verbele din **H** sunt sinonime (au descrieri semantice similare conform cadrelor VN) sau consecințe (starea sau acțiunea descrisă în **H** este un rezultat al evenimentului din **T**) ale verbelor din **T**, iar argumentele se potrivesc, sau dacă descrierea semantică a verbului corespunde cu o pereche element-atribut din **T**, ENTAILMENT

   3.b. Altfel, dacă verbele din **H** sunt antonime cu verbe din **T** sau mai generale decât acestea, iar argumentele se potrivesc, CONTRADICTION

   3.c. Altfel, UNKNOWN

Algoritmul descris mai sus folosește clasele Levin, așa cum sunt descrise în VN, pentru a rezolva inferențe textuale. Primul pas este acela de a atașa clasele Levin

corespunzătoare pentru fiecare verb din **T** şi **H**. Apoi încearcă extragerea de propoziţii atomice echivalente pe baza claselor Levin. Dacă verbele astfel găsite se potrivesc la nivelul argumentelor şi adjuncţilor, soluţia este ENTAILMENT; dacă argumentele nu se potrivesc sau verbele sunt în aceeaşi clasă dar sunt antonime, soluţia este CONTRADICTION; în orice alt caz, rezultatul este UNKNOWN. Dacă nu apar potriviri la nivel de clase Levin, algoritmul încearcă extragerea de propoziţii atomice echivalente pe baza descrierii semantice a verbelor, iar apoi face paşi similari ca mai sus. Dacă există verbe în **H** care nu au echivalent în **T** în urma paşilor descrişi mai sus, rezultatul este UNKNOWN.

4. Analiza corpusului

Aşa cum am afirmat şi în introducere, una dintre metodele cheie pentru determinarea inferenţelor textuale, conform intuiţiei noastre, este analiza verbelor. Scopul acestei secţiuni este acela de a prezenta studiul de fezabilitate efectuat pentru a determina utilitatea claselor Levin pentru problema inferenţelor textuale.

Testul de fezabilitate a fost realizat asupra unui corpus semnificativ, de 200 de perechi de inferenţă, extrase din setul de test RTE-5 (acesta este cel mai recent corpus de inferenţe textuale pentru limba engleză). Setul de perechi ales este tipic pentru datele RTE în ceea ce priveşte distribuţia perechilor: dintre cele 200 de perechi analizate, 100 (50%) sunt ENTAILMENT, 31 (15.5%) sunt CONTRADICTION şi 69 (35.5%) sunt UNKNOWN.

În urma analizei, fiecare pereche a fost clasificată în una dintre categoriile de mai jos:

- **VN (VerbNet)**, ceea ce înseamnă că aplicarea algoritmului definit în secţiunea 3 conduce la determinarea soluţiei corecte pentru perechea dată, în urma faptului că verbele din **T** şi **H** au aceeaşi clasă Levin, au descrieri semantice similare, sunt consecinţe pentru verbe din **T** sau pe baza compatibilităţii structurii de argumente;

- **NA (NeAtaşate)**, ceea ce înseamnă că algoritmul din secţiunea 3 nu poate găsi un verb în **T** care să aibă aceeaşi structură de argumente ca un verb din **H**, sau nu există verbe în **T** care să fie conectate semantic de verbe din **H**;

- **L (Lipsă)**, ceea ce înseamnă că unul dintre conceptele cheie din **H** nu apare în **T**;

- **AS (Alte Surse)**, ceea ce înseamnă că rezultatul corect poate fi dedus prin alte mijloace decât cele date mai sus (de exemplu rezoluţia anaforei sau cunoaştere ontologică). Este important de menţionat că, în urma analizei noastre, cele mai multe dintre perechile care cad în această categorie pot fi reduse la cazuri de comparare de seturi de trăsături pentru argumente şi adjuncţi.

Tabelul 1: Distribuţia soluţiilor pentru studiul de fezabilitate

Categoria	**Număr**	**Procent**
VN	76	38%
AS	59	29.5%
L	43	21.5%
NA	22	11%

După cum se poate vedea din distribuția soluțiilor pentru perechile analizate, cel mai mare câștig vine din folosirea VerbNet, urmat de alte surse de inferență (aproape toate aceste cazuri sunt instanțe de comparare de seturi de proprietăți). Distribuția susține intuiția noastră, întrucât toate cazurile de **NA** pot fi reduse la predicate care nu sunt echivalente, iar toate cazurile de **L** pot fi reduse la seturi de proprietăți care nu se potrivesc.

4.1 Perechi rezolvate folosind VN

Cea mai simplă utilizare a VN este alinierea verbelor din cadrul aceleiași clase Levin, așa cum se vede în exemplul 1.

Ex. 1: Potrivire exactă peste clase VN

***T**: MADAGASCAR'S constitutional court declared Andry Rajoelina as the new president of the vast Indian Ocean island today...(Curtea constituțională din Madagascar l-a declarat pe Andry Rajoelina președinte al marii insule din oceanul Indian)*

***H**: Andry Rajoelina was proclaimed president of Madagascar.(Andry Rajoelina a fost proclamat presedinte al Madagascarului.)*

Verbul *proclaim* din ipoteză face parte din clasa *say-37.7-1*, care conține și verbul *declare* din text. Aceste două verbe se potrivesc din punct de vedere al argumentelor (Andry Rojelina este *Theme*) și al adjuncților (*president of Madagascar*). Soluția pentru această pereche este ENTAILMENT, conform pasului **2.a.** din.

Folosirea descrierii sintactice a unui cadru și a descompunerii semantice atașate este descrisă în exemplul 2.

Ex. 2: Descriere sintactică și descompunere semantică

***T**: A court in Venezuela has jailed nine former police officers for their role in the deaths of 19 people during demonstrations in 2002. ...(Un tribunal din Venezuela a încarcerat nouă foști polițiști pentru rolul lor în morțile a 19 persoane în timpul demonstrațiilor din 2002)*

***H**: Nine police officers have had a role in the death of 19 people.(Nouă polițiști au avut un rol în moartea a 19 persoane.)*

Verbul *have* face parte din clasa *own-100*, care este descrisă sintactic *Theme1 V Theme2* și semantic *has_possession(E, Theme1, Theme2)*. În acest exemplu, *Theme1* este „*nine police officers*" iar *Theme2* este „*a role in the death of 19 people*"; așadar, situația descrisă în text („*their role*", care poate fi extins la „*the policeman's role*" prin rezoluția anaforei) susține ipoteza și duce la ENTAILMENT, conform pasului **3.a.** din algoritm.

Din cauza modului în care sunt definite clasele Levin, este posibil ca membrii aceleiași clase să nu fie sinonime, ci, mai mult, să fie antonime, așa cum se poate vedea în exemplul 3.

Ex. 3: Antonimie în aceeași clasă VN

***T**: BEIJING — China has rejected Coca-Cola Co.'s $2.3 billion bid to buy a major Chinese juice producer...(China a refuzat oferta de 2.3 miliarde de dolari a Coca-Cola Co.pentru cumpărarea unui mare producator chinez de băuturi răcoritoare)*

***H**: Coca-Cola buys Huiyuan Juice Group. (Coca-Cola cumpără Huinan Juice Group)*

Pentru a putea rezolva corect această pereche, verbele *reject* și *buy* din text trebuie relaționate. Verbul *reject* face parte din clasa *approve-77*; cazul întâlnit în **T** este descris sintactic de cadrul *Agent V Proposition* și semantic de *approve(during(E), Agent, Proposition)*. Deoarece verbul în discuție este antonim cu predicatul semantic folosit pentru descrierea sa, descrierea semantică devine *not(approve (during(E), Agent, Proposition))*. Clasa atașată verbului *buy* din **H** este aceeași cu cea din **T**. Pentru rezolvarea corectă a acestui punct, algoritmul trebuie augumentat cu un set de reguli de compunere de verbe, care să conțină reguli de tipul *not(approve(X))→not(X)*. Acest set de reguli poate fi creat relativ ușor, dat numărul de doar 145 de predicate booleene. Compunerea celor două predicate, *reject* și *buy*, duc la rezultatul corect, CONTRADICTION, conform pasului **3.b.** din algoritm.

În cazul în care nu există legătură între verbe pe baza claselor Levin, descrierea semantică a verbelor poate fi folosită pentru a determina o relație semantică, ca în exemplul 4.

Ex. 4: Potrivire pe baza descompunerii semantice

***T**: ... Fiat, the Italian car company that wants to acquire a stake in Chrysler. ...(Fiat, compania de automobile italiană care vrea să cumpere o parte din Chrysler)*

***H**: Fiat wants to gain possession of a stake in Chrysler.(Fiat vrea să câștige posesia unei părți din Chrysler)*

Deși nu există nici un verb în **T** care să fie în aceeași clasă cu *gain* (*get-13.5.1*), descrierea semantică a acestei clase, *has_possession(start(E), ?Source, Theme) transfer(during(E), Theme) has_possession(end(E), Agent, Theme) cause(Agent, E)*, este identică cu descrierea semantică pentru clasa *obtain-13.5.2-1*, care conține verbul acquire din **T**. Această potrivire, împreună cu potrivirea structurilor de argumente, duce la ENTAILMENT, conform pasului **3.a.** din algoritm.

VerbNet permite stabilirea de relații sintactice între verbe pe baza descrierii semantice a fiecărui cadru, după cum se observă în exemplul 5.

Ex. 5: Stabilirea de relații de consecință

***T**: ... Indira Gandhi, the former prime minister who was assassinated in 1984. ...(Indira Ghandi, fostul prim ministru care a fost asasinat în 1984.)*

***H**: Indira Gandhi died doing aerobatics.(Indira Ghandi a murit făcând acrobații)*

Descrierea semantică pentru clasa *murder-42.1*, din care face parte verbul *assassinate* este *cause(Agent, E) alive(start(E), Patient) not(alive(result(E), Patient))*, iar pentru clasa *disappearance-48.2*, cea a verbului *die*, este *disappear(during(E), Theme)*. Deoarece *disappear* este sinonim cu *die* (*not alive*), iar argumentele celor două descrieri semantice se potrivesc (*Theme* și *Patient* sunt similare), se poate deduce că verbul *die* este o consecință a verbului *kill*. Rezultatul este CONTRADICTION, conform pasului **3.b.** din algoritm, deoarece verbul din **H** are un adjunct (*doing aerobatics*) în contradicție cu adjuncții din **T**.

4.2 Perechi rezolvate folosind L și NA

În cazul problemei de rezolvare a inferențelor textuale cu trei soluții posibile (ENTAILMENT, CONTRADICTION, UNKNOWN), una dintre principalele dificultăți este aceea de a determina cazurile UNKNOWN. Conform intuiției pe baza căreia am făcut această analiză, cazurile UNKNOWN apar, de cele mai multe ori, din două motive: una sau mai multe din entitățile din **H** nu apare în **T** (**L**) sau toate entitățile din **H** apar în **T** dar nu sunt conectate de un același predicat (**NA**). Un caz de entități din **H** care nu sunt conectate în **T** este dat în exemplul 6.

Ex. 6: Entități din **H** care nu sunt conectate în **T**

> ***T****: Currently, there is no specific treatment available against dengue fever, which is the most widespread tropical disease after malaria. ... the mosquitoes that transmit dengue ...(În acest moment nu există un tratament pentru febra dengue, care este cea mai răspândită boală tropicală după malarie...țânțarii care transmit dengue)*
>
> ***H****: Malaria is the most wide-spread disease transmitted by mosquitoes. (Malaria este cea mai răspândită boală tramisă de țânțari)*

În cadrul textului, cuvântul *malaria* apare doar o singură dată, și nu este conectat cu noțiunea de *mosquitoes*; conform pasului **2.c.** din algoritm, acest caz este UNKNOWN.

În urma analizei, toate cele 22 de cazuri de **NA** și cele 43 de cazuri de **L** sunt UKNOWN, și constituie 65 din totalul de 69 de astfel de cazuri. Celelalte 4 cazuri se datorează faptului că verbele din **H** nu se regăsesc în **T**, după cum se observă în ex. 7.

Ex. 7: Determinarea de UNKNOWN folosind nepotrivirea verbelor

> ***T****: ...The iron had to be heated to a precise cherry red color and beaten by the right combination of hammer blows. Mediocre work could hide problems.(Fierul trebuia încălzit până ajungea la culoarea corectă și trebuia bătut cu o serie de lovituri cu ciocanul speciale. Munca mediocră putea ascunde probleme.)*
>
> ***H****: Titanic sank in 1912. (Titaniicul s- scufundat în 1912.)*

Clasa Levin pentru verbul *sink* este *other_cos-45.4*, și are descrierea semantică *state(result(E), Endstate, Patient)*. Singurul verb din **T** similar semantic este *heat*, care face parte din aceeași clasă ca și verbul *sink*, dar nu are ca argument *Titanic* sau ca adjunct *1912*, fapt ce duce la soluția UNKNOWN, conform pasului **3.c.** din algoritm.

4.3 Perechi rezolvate folosind AS

După cum am menționat mai sus, **AS** se referă la rezolvarea inferențelor textuale prin metode diferite de compararea de verbe (cel mai frecvent prin compararea de proprietăți ale entităților din **T** și **H**). Dificultatea acestei abordări provine clasificarea atributelor; soluția în acest caz este utilizarea unei ontologii (testele preliminare au demonstrat că atât WordNet (Fellbaum, 1998) cât și SUMO (Niles, Pease, 2001) sunt utile în acest sens). Un exemplu de utilizare a cunoașterii ontologice pentru a putea rezolva inferențe textuale este dat în exemplul 8.

Ex. 8: Rezolvarea inferențelor textuale folosind cunoaștere ontologică

> ***T****: ... Mr. Barnes was an offensive lineman in the old American Football League, ... (Domnul Barnes a fost atacant în vechea liga de fotbal american)*

***H**: Ernie Barnes was an athlete. (Ernie Barnes a fost atlet)*

În acest caz, SUMO conţine informaţia că „lineman" este un tip de atlet, fapt ce duce la soluţia ENTAILMENT pentru această pereche; de asemenea, verbele copulative trebuie văzute ca o modalitate de asignare de proprietăţi.

5. Concluzii

În această lucrare am descris o nouă metodă pentru rezolvarea inferenţelor textuale folosind semantică predicaţională şi structuri de trăsături. Pentru a dovedi fezabilitatea acestei metode am realizat un studiu detaliat a 200 de perechi de inferenţe din setul de test din RTE-5. Pentru determinarea semanticii predicaţionale am examinat utilitatea VerbNet, cu ajutorul căruia pot fi rezolvate direct 38%. De asemenea, am dat şi o primă variantă a unui algoritm de rezolvare a inferenţelor textuale folosind VerbNet. Un alt avantaj al VN este acela că este aliniat cu o serie de resurse ca FrameNet şi WordNet, ceea ce înseamnă că transpunerea sa în limba română poate fi făcută relativ uşor (principalele dificultăţi sunt legate de alinierea cadrelor sintactice din română cu cele din engleză). Dată o variantă de VerbNet pentru limba română, algoritmul propus în această lucrare ar putea rezolva direct perechi de inferenţe în limba română.

Întrucât utilitatea VerbNet, în conjuncţie cu un context de aplicare (un set de reguli), pentru rezolvarea de inferenţe a fost dovedită, iar utilitatea ontologiilor pentru extragerea de seturi de atribute a fost arătată, scopul în acest moment este implementarea acestor idei într-un sistem funcţional. De asemenea, urmează să dezvoltăm o metodă pentru transpunerea VerbNet-ului în limba română pe baza alinierii cu WordNet.

Referinţe bibliografice

Baker, C. F. Fillmore, C. J., Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics* (COLING/ACL-98), pages 86–90, Montreal.

Burchardt, A., Frank, A. (2006). Approximating Textual Entailment with LFG and FrameNet Frames. *Proceedings of the 2nd PASCAL Workshop on Recognising Textual Entailment*, Venice, Italy.

Dagan, I., Glickman, O. (2004). Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining*, Grenoble, France.

Dagan, I. Magnini, B. Glickman, O. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenge Workshop for Recognising Textual Entailment*, pages 1–9, 11–13 April, 2005, Southampton, U.K.

Fellbaum, C. editor. (1998). WordNet: An Eletronic Lexical Database. *Language, Speech and Communications*. MIT Press, Cambridge, Massa-chusetts.

Hickl, A., Bensley, J. (2007). A Discourse Commitment-Based Framework for Recognising Textual Entailment. In *Proceedings of the ACL-PASCAL Workshop*

on Textual Entailment and Paraphrasing. Pages 185-190. 28-29 June, Prague, Czech Republic.

Hickl, A. (2008). Using Discourse Commitments to Recognize Textual Entailment. *Proceedings of the 22nd Conference on Computational Linguistics (COLING 2008)*. Manchester, United Kingdom.

Kipper, K., Dang H. T., Palmer, M. (2000). Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. *Ph.D. thesis, Computer and Information Science Dept.*, University of Pennsylvania, Philadelphia, PA, June

Kipper-Schuler, K, Korhonen, A., Ryant, N., Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. *Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy. June, 2006.

Korhonen, A., Briscoe, T. (2004). Extended Lexical-Semantic Classification of English Verbs . In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, Boston, MA

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation, *University of Chicago Press*, Chicago, IL.

Niles, I., Pease, A. (2001). Towards a standard upper ontology. In Proceedings *of the 2nd International Conference on Formal Ontology in Information Systems* FOIS-2001.

Moens, M., Steedman, M. (1988). Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14:15–38.

Palmer, M., Gildea, M., Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles, Computational Linguistics, 31:1 , pp. 71-105, March, 2005.

SUMMARIES/ABSTRACTS

METHODOLOGY FOR ESTABLISHING AND ANALYZING ANNOTATED SPEECH CORPUSES – THE CASE OF THE CORPUS SROL

HORIA-NICOLAI TEODORESCU

Institute for Computer Science of Romanian Academy, Iaşi, România
Technical University of Iaşi, Faculty of Electronics, Telecommunications and Information Technology, Iaşi – România

hteodor@etti.tuiasi.ro

Abstract

The purpose of the paper is to thoroughly describe a comprehensive methodology for the establishment and analysis of a spoken language corpus, with reference to the Romanian language. While it is of pragmatic and methodological essence, the paper also aims to draw the attention that previously developed corpuses that were used in linguistic inference are unsatisfactorily from the current requirements, hence leading to erroneous conclusions regarding some of the characteristics of spoken languages.

NEW TECHNIQUES FOR THE IDENTIFICATION OF VOWEL AREAS UTTERED IN ROMANIAN LANGUAGE

M. ZBANCIOC[1,2], H.N. TEODORESCU[1, 2], M. FERARU[1]

[1]*Institute for Computer Science of Romanian Academy, Iaşi, România*
[2]*Technical University of Iaşi, Faculty of Electronics, Telecommunications and Information Technology, Iaşi – România*

{hteodor, zmarius}@etti.tuiasi.ro

Abstract

A new set of segmentation techniques is presented, techniques used in order to identify vocalic areas, information that will subsequently be used by the extraction instruments for the fundamental frequency F_0, moreover for determining the value of the formants F_1,...,F_4. The segmentation phase is important because its errors will have a direct effect on the performances of the pitch extractor. The segmentation precision of our instrument is compared to that of the Praat tool, using high precision annotated files. In order to reduce the running time, a series of optimizations were made to the computation of the autocorrelation function, by applying several recurrent algorithms.

METHODOLOGICAL ASPECTS OF ORGANIZING DATA AND STATISTICAL ANALYSIS OF EMOTIONAL VOICE

HORIA-NICOLAI TEODORESCU[1, 2], IOAN PĂVĂLOI[1], MONICA FERARU[1]

[1]*Institute for Computer Science of Romanian Academy, Iași, România*
[2]*Technical University of Iași, Faculty of Electronics, Telecommunications and Information Technology, Iași – România*

{hteodor}@etti.tuiasi.ro

Abstract

We present a methodology and a program for statistical analysis of emotional voices. Moreover, we present some preliminary results regarding the organization of data in an application for voice signal files. The program allows statistical analysis of the vowels and semivowels of formantic characteristics on the subclasses of data files selected according to user characteristics. The application allows a degree of refinement in the voice sounds analysis.

ROMANIAN INTONATIONAL CONTOUR EDITOR BASED ON FUNCTIONAL PROSODIC FORMS

VASILE APOPEI, DOINA JITCĂ, OTILIA PĂDURARU

Institute of Computer Science,
Romanian Academy – Iasi Branch

vapopei@ iit.tuiasi.ro, jdoina@iit.tuiasi.ro

Abstract

The paper presents an intonational tree editor that may be used in Romanian intonation analysis and synthesis. The editor helps the user in spanning the input phrase into multi level prosodic units in order to generate a certain intonational contour in phrase utterance synthesis. In essence the tool offers commands for defining the tree structure and the unit-objects from the nodes and the leaves. The most important attribute in unit-object defining is a functional one at communication act level that assigns it a certain F0 contour pattern (an elementary melodic contour). By using this editor, the understanding of the intonation model is improved. The section 2 presents the intonation model from its functional and hierarchical perspective. The editor functionalities are presented in section 3 and they refer to the intonational tree file management, to the intonational tree editing and to their conversion into speech outputs corresponding to synthesized input phrase utterances.

SUMMARIES/ABSTRACTS

CORPUS FOR GNATHOPHONY: PROTOCOL, METHODOLOGY, ANNOTATION

HORIA-NICOLAI TEODORESCU[1,2], ALINA UNTU[1]

[1]Technical University of Iasi, Faculty of Electronics, Telecommunications and Information Technology, Iași - Romania
[2]Institute for Computer Science of the Romanian Academy, Iași - Romania;

{hteodor, auntu}@etti.tuiasi.ro

Abstract

We present aspects regarding the extension of a gnathophonic sounds micro-corpus, examples of formantic and temporal analysis applied to the fricative consonants in prosthesis and edentulous cases, as well as preliminary results.

ELECTRONIC DEVICE FOR PARALLEL PREPROCESSING OF THE VOCAL SIGNAL SPECTRUM

HULEA MIRCEA[1], UNTU ALINA[2]

[1]Technical University „Gheorghe Asachi" of Iasi, Faculty of Automatic Control and Computer Science, Iași – Romania;
[2] Technical University „Gheorghe Asachi", Faculty of Electronics, Telecomunications and Information Technology, Iași – România;

mhulea@tuiasi.ro, auntu@etti.tuiasi.ro

Abstract

This paper presents a hardware device used for stimulation of an analogue neural network able to perform speech recognition. Considering that spiking neurons are coincidence detectors, the recognition of phonemes should be performed by detection of concurrent events generated by audio signal. Considering that the vowels are characterized by a number of frequencies which are summed during their reception, the system splits the vocal spectrum of the audio signal into frequency channels. The main advantage of this prototype is that it is adapted to generate the trains of impulses needed by a network of spiking neurons to perform detection of singular vowels and vowel combinations. To increase the vowel recognition accuracy the central frequencies of the filters are chosen taking into account the vowels formants. Due to the fact that the device could be used only on vowel detection, as a future research goal we want to test these operation principles in discrimination of the frequency channels activated during consonants reception.

CONTINUOUS FLUX ROMANIAN LINGUISTIC RESOURCES

DAN CRISTEA

Institute of Computer Science, Romanian Academy, Iași – România;
Faculty of Computer Science, „Al. I. Cuza" University, Iași - România

dcristea@info.uaic.ro

Abstract

The paper raises the issue of a legislative initiative for acquiring large scale language resources. It militates for a large awareness campaign that would bring to stage the importance of storing and preservation for research purpose, in electronic form, of all textual documents which go to print, dayly, in Romania. The paper refines the technological steps for this achievement, and stresses legal aspects that should be taken into consideration.

ELECTRONIC COMMUNICATION AND OUR ORTOGRAPHICAL PROBLEMS - WITHOUT SOLUTIONS?

LUCIAN CHIȘU

National Museum of Romanian Literature, Bucharest

lucianchisu@gmail.com

Abstract

This text is a continuation of the author's intervention, published under the title Electronic communication and our ortographical problems (AUSH, series Journalism, 2003). At the time of his first intervention, the digital language in Romanian had a rather transitory aspect. The author considers that, for its optimal usage, two kinds of interventions would be necessary. On one hand, the intervention of technical and system engineers towards creating the software to adapt the keyboard to the Romanian language. On the other hand, the firm attitude of the specialists (liinguists, philologists, the cultural elite) towards respecting the ortographical and literary language norms in the case of the digital language, which is about to become the most consistent form of communication between institutions and individuals. The conclusion of the new research, including the 2003-2010 period, is the following: although software exists and Romanian language keyboards are commercialised, and although laws have been passed sanctioning the bad usage of the literary language, the deviations, mistakes and hesitations from the early days are still present in a rather worrying percentage. Official people websites are indicated, some belonging to the Ministry of Culture and Ministry of Education, where ortographical reglementations, far from being put to use, are avoided, in some cases with the express indication to elude diacritics.

FACTORS THAT INFLUENCE THE OPTIMAL MOMENT FOR MIGRATING TO THE CORRECT DIACRITICALS IN THE ROMANIAN LANGUAGE

BOGDAN STĂNCESCU

S.C. Moongate Video Production srl, București – România

bogdan@moongate.ro

Abstract

The author estimates that during 2010-2015 most Romanian language content available in digital form will migrate from using the legacy characters with diacritics (using cedilla, per ISO-8859-2) to the proper characters (using comma, per Unicode 3.0). The paper synthesizes the relevant factors that influence the ideal moment for migration, from case to case.

AUTOMATIC EXTRACTION OF MULTILINGUAL CORPUS

TIBERIU BOROȘ, DAN TUFIȘ, ALEXANDRU CEAUȘU

Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

{tibi,tufis,aceausu}@racai.ro

Abstract

The lack of resources is the primary impediment in technological progress in the automatic processing of most natural languages. The success of statistical approaches in automatic translation for wide spread languages has shown the importance of collecting and preprocessing large volumes of linguistic resources. Parallel corpuses and recently, comparable corpuses, represent the primary resource for cross and multilingual applications. This article presents a set of tools that can be used in automatic extraction of parallel and strongly comparable corpuses.

LEXICAL SEMANTICS MODELING AND ROBUST PARSING FOR ROMANIAN, FRENCH, AND GERMAN THESAURI

NECULAI CURTEANU[1], MIHAI ALEX MORUZ[1,2], DIANA TRANDABĂȚ[1,2]

[1]*Institute of Computer Science, Romanian Academy, Iași – România*;
[2]*Faculty of Computer Science, „Al. I. Cuza" University, Iași - România*

{curteanu, mmoruz, dtrandabat}@iit.tuiasi.ro

Abstract

This paper presents a lexical-semantics cross-linguistic analysis of the largest thesauri currently existing for Romanian, French, and German, and a new, robust and portable method for their entry parsing, based on the technique of Segmentation-Cohesion-Dependency (SCD) configurations. The general idea behind parsing a thesaurus or dictionary is to transform a raw text dictionary entry into an indexable linguistic database of lexical-semantics (sense) trees. The SCD parsing configurations are applied successively on a thesaurus entry in order to identify its lexicographic segments (the first SCD configuration), to extract the tree of its senses and subsenses (the second one), and to parse its atomic and non-atomic sense definitions (the third one). Using previous results on **DLR** (The Romanian Thesaurus – new format), the present paper adapts and applies the SCD-based technology to other four large and complex thesauri: **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), and **GWB** (Göthe-Wörterbuch). This experiment, which was illustrated on significantly large parsed entries of the mentioned thesauri, proved the main achievements of the parsing technique based on SCD-configurations: efficiency, robustness, portability. These qualities, when compared with the classical methods for dictionary entry parsing, derive mainly from at least two SCD-specific facts: the sense tree extraction is performed on the sense marker sequences exclusively, while the processes of sense tree extraction and (atomic) sense definition parsing are shown to be completely separable.

THE ACHIEVEMENT OF A ROMANIAN TREEBANK

CENEL-AUGUSTO PEREZ

Faculty of Letters, „Al. I. Cuza" University, Iași - România
Faculty of Computer Science, „Al. I. Cuza" University, Iași - România
cperez@info.uaic.ro

Abstract

We inventorize the outcomes of the achievement of the Romanian language syntactic corpus, stressing on the problems met during the acquisition of this corpus (we will also mention some steps in the creation of the corpus) and on the solutions of these problems.

MONITORING NEOLOGISMS IN NEWSPAPERS WITHIN THE NEOROM PROJECT

ANA-MARIA BARBU

„Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Bucharest, Romania

anamaria.barbu@g.unibuc.ro

Abstract

This paper describes the contribution of the Romanian research team within the international project NEOROM. The project aims at building a lexical database consisting in neologisms which occur in newspapers written in all Romance languages since 2004. The paper has three parts. In the first one, we present the monitoring system adopted by the Romanian team, in the second one we describe the NEOROM interface which allows to register the neological words in the database, whereas in the third part we give a short analysis of the words in the database from the viewpoint of their formation.

EMOTIONS IN WORDS: DEVELOPING A MULTILINGUAL RESOURCE

VICTORIA BOBICEV, VICTORIA MAXIM, TATIANA PRODAN,
NATALIA BURCIU, VICTORIA ANGHELUŞ

Technical University of Moldova, Chişinău, Republic of Moldova
vika@rol.md, maxivica@yahoo.com, tatiana.ursulenco@gmail.com,
natusicb@yahoo.com, lazu_vic@yahoo.com

Abstract

In this paper we describe the process of Russian and Romanian WordNet-Affect creation. WordNet-Affect is a lexical resource created on the basis of the Princeton WordNet which contains information about the emotions that the words convey. It is organized in six basic emotions: anger, disgust, fear, joy, sadness, surprise.

We translated the WordNet-Affect synsets into Russian and Romanian and created an aligned English – Romanian – Russian lexical resource. The resource is freely available for research purposes.

A TRAINABLE QA SYSTEM FOR ROMANIAN

DAN ȘTEFĂNESCU, RADU ION, ALEXANDRU CEAUȘU, DAN TUFIȘ, ELENA IRIMIA, VERGINICA BARBU-MITITELU

Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

{danstef, radu, aceausu, tufis, elena, vergi}@racai.ro

Abstract

This paper presents a Question-Answering system developed at the Research Institute for Artificial Intelligence (RACAI) in a context provided by a national project. The system was objectively evaluated by the organizers of the CLEF QA competition, under the ResPubliQA task for Romanian. We describe the combination of various relevant factors of the system used for the identification of the most relevant paragraphs as answers to natural language questions. The system is available online on the web page of RACAI. It is fully trainable and its functionality is independent of the genre characterizing the training data.

A MACHINE TRANSLATION SYSTEM FOR ROMANIAN AND FRENCH

MIRABELA NAVLEA, AMALIA TODIRAȘCU

LiLPa, Université de Strasbourg, Strasbourg – France, 22 rue René Descartes, BP 80010, 67084 Strasbourg cedex France;

mirabela.navlea@yahoo.fr, todiras@unistra.fr

Abstract

We present an ongoing project aiming at the development of a Machine Translation system for Romanian and French. We adopt a factored phrase-based statistical machine translation method and we combine several linguistic categories (word form, POS tag, morpho-syntactic properties). We use existing parallel corpora and we also build our own parallel, sentence and word aligned corpora. The corpora are tagged, lemmatized and annotated at chunk level. The paper focus on the presentation of aligned corpora. We present a detailed linguistic analysis of word alignment errors and we define a set of repairing rules, used to improve the quality of the word alignment.

INTEROPERABLE AND MULTILINGUAL WEB SERVICES

RADU ION, ALEXANDRU CEAUȘU, DAN ȘTEFĂNESCU, DAN TUFIȘ

Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

{radu, aceausu, danstef, tufis}@racai.ro

Abstract

Interoperability problems of the linguistic tools and resources are major concerns of current research in Natural Language Processing. Most of the new technologies based on (web) Service Oriented Architectures are important steps towards ensuring interoperability. When web services are language independent or are easily adaptable to new languages, interoperability and multilinguality criteria are largely satisfied. Orchestration of different web services is usually done by choosing a common format of the I/O data. In this respect, several platforms for integrating resources and tools developed in various programming languages and running under various operating systems already appeared. In this paper we present some of the linguistic web services of the Research Institute for Artificial Intelligence of the Romanian Academy (ICIA) adapted to the web services platform WebLicht. This platform is an environment where web services can interact because of the fact that their I/O parameters are converted and required to be in TCF (Text Corpus Format).

HYPONYMY PATTERNS FOR ROMANIAN

VERGINICA BARBU MITITELU

Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

vergi@racai.ro

Abstract

Hyponymy is a lexical-syntactic relation that was not rigorously studied in Romanian linguistics. Yet, it offers engineers a very efficient way of organizing the lexical material useful in the numerous tasks that they develop and that imply natural language processing. We present here two ways of identifying hyponymy patterns in Romanian, the results obtained, and their evaluation. We foresee uses of these patterns in the end of the article.

GENERATIVE MECHANISMS OF DERIVATIONAL MORPHOLOGY

MIRCEA PETIC

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, Chisinau, Republic of Moldova

mirsha@math.md

Abstract

The article studies the problems of a derivative generator development. The starting point is a lexicon that provides storage of derivatives in a lexicon, which contains not only the graphical representation of derivatives but also their constituent morphemes. This allows studying and formulating rules that would generate derivatives taking into account some restrictions.

CREATING A SEMANTIC ROLE PARSER FOR ROMANIAN

DIANA TRANDABĂȚ, DAN CRISTEA

Institute of Computer Science, Romanian Academy, Iași, Romania
Faculty of Computer Science, „Al. I. Cuza" University Iași, Romania

{dtrandabat, dcristea}@info.uaic.ro

Abstract

Semantic parsing, by identifying and classifying semantic entities in context, as well as the relations between them, has a great potential for applications such as text summarization, question-answering or machine translation. Thus, by developing a system that automatically annotates semantic for the Romanian language, this paper represents an important intermediate step towards automatic natural language understanding. For the creation of the semantic role parser for Romanian, it was first necessary to develop a training corpus, annotated with semantic roles. This annotated resource of semantic roles for the Romanian language was produced using as starting point a resource developed for English (FrameNet), and applying an automatic transfer method. Subsequently, using a platform for developing supervised semantic role labeling systems (PASRL), several learning algorithms were trained on the developed corpus and the best obtained model was saved. We discuss the results of applying this technique for Romanian, with the conviction that other languages could also benefit from using the same approach.

INTERNET SURVEILLANCE: CONSUMERS OPINION ABOUT CERTAIN PRODUCTS OR EVENTS

ADRIAN IFTENE, ALINA-ELENA MIHĂILĂ, GEORGE-ALEXANDRU VLAD, GETA STANCU

Faculty of Computer Science, „Al. I. Cuza" University of Iasi, Romania

{adiftene, elena.mihaila, george.vlad, geta.stancu}@info.uaic.ro

Abstract

This paper approaches and offers a solution to a modern day concern: relevant information retrieval and surveillance through the internet. The increasing volume of information that exists on web sites, forums or personal web pages, makes the process of searching for information that is relevant for us to become very complex and time consuming. In addition to their success, social networks like Twitter, MySpace, Facebook, Flickr have more and more users with common interests, and their gathered information has an increasing volume. In such social networks starting from a topic, users can express freely their opinions, add links or relevant photos. The system that we built uses information collected from the Internet and offers to users an easier way to find out positive and negative opinions about a topic. Information is searched on web pages (we prefer blogs, forums or users comments) and on Twitter. Identifying and classifying opinions is done by identifying some emotional triggers and by calculating some valences related to the context in which they appear.

USING VERBS FOR DETERMINING TEXTUAL ENTAILMENT

MIHAI ALEX MORUZ[1,2]

[1]*„Al.I.Cuza" University, Faculty of Computer Science, Iasi – Romania*
[2]*Romanian Academy, Iasi Branch, Institute for Computer Science – Romania*

mmoruz@info.uaic.ro

Abstract

This paper describes a method of using verbs and their argument structure for solving Textual Entailment for English and Romanian. The analysis carried out over a set of entailment pairs from the RTE-5 test set proves that predicational semantics, as defined by Charles Fillmore and Beth Levin, is useful for solving textual entailment, as 38% of the pairs we analyzed are directly solved by using this type of semantics and 11% of the pairs are solved by using verbal analysis.

INDEX DE AUTORI

Apopei Vasile	45
Angheluș Victoria	141
Bobicev Victoria	141
Barbu Ana-Maria	131
Barbu-Mititelu Verginica	153, 185
Boroș Tiberiu	103
Burciu Natalia	141
Ceaușu Alexandru	103, 153, 175
Chișu Lucian	81
Cristea Dan	73, 203
Curteanu Neculai	113
Feraru Monica	23, 35
Hulea Mircea	61
Iftene Adrian	213
Ion Radu	153, 175
Irimia Elena	153
Jitcă Doina	45
Maxim Victoria	141
Mihăilă Alina-Elena	213
Moruz Mihai Alex	113, 223
Navlea Mirabela	165
Păduraru Otilia	45
Păvăloi Ioan	35
Perez Cenel-Augusto	123
Petic Mircea	195
Prodan Tatiana	141
Stancu Geta	213
Stăncescu Bogdan	89
Ștefănescu Dan	153, 175
Teodorescu Horia-Nicolai	13, 23, 35, 51
Todirașcu Amalia	165
Trandabăț Diana	113, 203
Tufiș Dan	103, 153, 175
Untu Alina	51, 61
Vlad George-Alexandru	213
Zbancioc Marius-Dan	23